Essays on Scale and Scope Economies in the English NHS

This Thesis is Submitted
In Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

The Division of Health Research
Lancaster University

Robert J. Willans

July 2023

Abstract

Title: Essays on Scale and Scope Economies in the English NHS
Author: Robert Willans
Degree: Doctor of Philosophy
Submission Date: July 2023

This thesis considers the relationship between the size and structure of hospital services and their costs. It is often assumed that service amalgamation ought to yield lower costs through economies of scale. However, empirical evidence for this is limited and often dated. This thesis is structured as a series of essays evaluating scale and scope economies using parametric methods applied to cost and activity data from the English National Health Service, covering April 2013 - March 2019. The first three empirical chapters consider the relationship between size and average healthcare cost, whilst the last explores how the configuration of services affects the cost of hospital healthcare provision. Several parametric specifications and methods are used to evaluate scale economies using the dataset. Results show small but positive economies of scale for various specifications up until around 1,000-1,200 beds, which constituted most hospitals in the sample. Scale economies after this point varied according to the method used, suggesting that methodology, particularly the choice of the functional form, may partly explain variation in the literature. Differences are observed between the direct estimation of a long-run cost function and a long-run function obtained from the envelope of short-run functions. Scale economies were also lower in London and surrounding areas due to higher wage rates for non-medical staff. The analysis of scope economies found that general surgery demonstrated the highest degree of scope economies compared to all other outputs, with general medicine and obstetrics/gynaecology also exhibiting positive scope economies to a lesser degree. General surgery, general medicine and obstetrics/gynaecology may benefit from lower costs when collocated with other activities. This thesis updates prior estimates of hospital economies of scale using rich data. It provides insights into methodological sources of variation, leading to conclusions of interest for policy in planning hospital service provision and the effects of scale and scope.

Table of Contents

List of Tables

List of Figures

Acknowledgements

This thesis would have of course not been possible without the help, support, and guidance of many people.

I would like to thank my supervisory team, Katharina Janke, Bruce Hollingsworth, and Mike Tsionas, who have offered unstinting assistance and encouragement, making suggestions on drafts and offering advice throughout the process. As a part time distance learning student, finding time and engaging with academia at a distance has occasionally been challenging, however the support I have received has enabled me to greatly improve as a researcher.

The availability of data being crucial for this thesis, I would also like to thank those who helped facilitate access to datasets, especially where these needed additional processing, particularly the costing team at NHS England/Improvement and the workforce analysis team at NHS Digital who often went above and beyond the normal course of duty.

Family and friends have also been a great help, offering sounding boards, proof reading and a welcome balance from writing when needed.

Lastly, and most importantly, I owe a particular debt of gratitude to Lucy for providing incredible support and encouragement as needed, and for John and Alice, born whilst this PhD was completed.

Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. An earlier version of Chapter 2 was presented at the EuHEA PhD conference in September 2021.

Chapter 1

Introduction: Firm Theory and Multiproduct Hospitals

1.1 Thesis Structure and Overview

This thesis is structured as a series of essays evaluating scale and scope economies in English hospitals, using parametric methods and data from April 2013 - March 2019. The thesis as a whole aims to investigate the presence of, and limits to, scale and scope economies in the provision of hospital healthcare.

Chapter 2 uses a multiproduct translog functional form to directly estimate a long-run cost function and calculate scale economies using a multiproduct scale economy index. The main result finds small but positive scale economies up to 1,200 beds, higher than previous estimates, with constant scale economies after that point. The median organisation had a scale economy index of 1.04, indicating that a 1% increase in the cost of inputs was associated with a 1.04% increase in outputs, with smaller organisations having higher economies of scale. It was also noteworthy that hospitals in London recorded lower economies of scale due to increased labour costs. Higher wages in these areas with similar labour productivity means that expansion does not give the same benefits as elsewhere.

Chapter 3 repeats the estimation, but this time using a quadratic functional form, contrasting the results and evaluating whether there are any theoretical reasons to prefer a particular functional form. Results for the quadratic form of the cost function generated a distribution of scale economy estimates. However, the relationship to size was less pronounced. Small positive scale economies were noted amongst English hospitals but did not appear to have a strong association with size measured by beds. Lower scale economies were again observed for hospitals in London and surrounding areas. As identical data and variables were used for the two chapters, differences may be due to the functional form of the cost function. The choice of functional form may be a source of variation in results from studies that use only one functional form.

Chapter 4 estimates scale economies via an indirect method to account for the fact that observed values may not be on the frontier of the cost function. Hospitals may not have optimal capital for their level of output and may not be producing at the lowest cost. This chapter first estimates a short-run cost function contingent on capital levels proxied by beds. An envelope of the short-run cost functions is calculated to estimate a cost function and associated scale economies where capital would be at the optimal level. Results returned higher estimates of scale economies, as might be expected from the underlying theory. However, optimal capital values, as computed, were sensitive to different functional forms and specifications, returning implausible values, particularly for the quadratic functional form. Variation in computed optimal capital values may represent weaknesses in using beds as a proxy value of capital and a lack of convexity, meaning the partial derivatives of cost with respect to beds may not reflect global minima. Future studies attempting to correct for non-optimal current capital levels may need to demonstrate that bed numbers or other proxy variables are appropriate capital measures and that calculated optimal values are plausible.

Chapter 5 briefly explores scope economies, using clinical specialties as outputs. A cost function is estimated using activity aggregated by clinical specialties rather than care settings. This output classification is a more natural subdivision of activity which corresponds better to the natural construction of hospital departments or services. Scope economies are calculated according to the cost difference between joint and separate production, with a subsequent discussion of policy implications. Of the specialty groups, general surgery demonstrated the highest degree of scope economies compared to all other outputs, with general medicine and obstetrics/gynaecology exhibiting positive scope economies to a lesser degree. Diagnostic and imaging services and orthopaedics did not exhibit notable scope economies. This chapter provides evidence that specialties demonstrating scope economies are cheaper to provide as part of a suite of hospital services and may, therefore, not be amenable to centralisation in fewer units. Conversely, those specialities which did not demonstrate scope economies, like diagnostic services, may be better candidates for rationalisation in fewer providers.

The remainder of this chapter identifies some of the theoretical and measurement issues with estimating the cost functions and deriving scale economy estimates. I note common issues in the literature and datasets used in subsequent essays and document the reasoning behind particular methodological choices. Many of the issues identified stem from the complexity and heterogeneity of hospital inputs and outputs. Hospital care can cover many different medical specialties and sub-specialties

- indeed, hospitals are where almost all specialist care is delivered. The interplay of differing inputs and outputs in a single organisation means that hospital cost analyses inevitably have to deal with the multiproduct environment. This multiproduct environment complicates analyses beyond the canonical textbook case of single-product organisations. Costs are no longer allocable to individual products, as joint production means that at least some costs cannot be allocated to individual outputs. The effect of scale on output is also conceptually different in the multiproduct case, as it is necessary to consider *which* outputs are considered and what is meant by ‘scale’. Measurement issues also arise, as the heterogeneous treatments offered in hospitals need to be reasonably represented in cost models.

Below, some of the theoretical concepts used in subsequent chapters are set out, together with short discussions of common measurement and estimation issues for multiproduct hospital cost functions. Firstly, the basic case of single-product cost-minimisation is defined in section 1.2, followed by how this changes in the multiproduct case in section 1.3. The definition of multiproduct scale and scope economies are discussed in section 1.3.3 and section 1.3.4 respectively. For parametric analysis, the functional form of the cost equation is important. Some common functional forms for cost equations are discussed in section 1.5.3. Measurement issues are discussed in section 1.4, including definitions of a hospital, conception and measurement of hospital outputs, and issues of casemix and coding. This measurement section also deals with definitions of size and how well size can be measured. Finally, estimation methods and other considerations are discussed in section 1.5, covering parametric and non-parametric techniques and model specifications.

1.2 Single-Output Production

1.2.1 Definitions

A cost function $c(w, y)$ expresses the *minimum cost* incurred by a firm producing output y under a given technology, facing input prices $w = (w_1, ..., w_n)$. As an accounting identity, total costs are equal to the sum of all inputs q_i multiplied by their respective input prices w_i, $\sum_i q_i w_i$. For cost-minimising organisations, demand for inputs is a function of the desired level of output y. The relationship between outputs and inputs is specified in the production function, dictated by the technology used. Therefore, the cost function can be expressed in terms of output y, input prices w as $C(y, w)$ for a given production technology.

From the cost function, several useful concepts can be derived. Firstly, Shephard's lemma gives us the cost-minimising input quantities for each level of output:

$$\frac{\partial C(w,y)}{\partial w_i} = x_i(w,y) \tag{1.1}$$

Secondly, differentiating with respect to output gives a marginal cost expression, evaluating the incremental increase (or decrease) in total cost caused by an incremental increase (decrease) in output:

$$MC = \frac{dC(w,y)}{dy} \tag{1.2}$$

Finally, when cost is plotted against output, the marginal cost is the slope of the cost curve at that point. The average cost of output at a given point can be defined as the slope of a ray extending from the origin. Average costs are consequently total costs C divided by total output y:

$$AC = \frac{C(w,y)}{y} \tag{1.3}$$

1.2.2 Isocosts and Isoquants

Costs incurred by an organisation can be represented using isocost lines and outputs using isoquant lines. Isocost lines show combinations of inputs yielding identical total costs at the set input prices. Isoquant lines show identical output levels at each bundle of inputs. Production decisions can be described using the relationship of isocost and isoquant lines.

The two input case is illustrated in Figure 1.1. X and Y axes represent the quantity of inputs q_1 and q_2 used. Equivalent costs are represented by linear isocost lines $C = q_1 \cdot w_1 + q_2 \cdot w_2$. Each isocost line has slope dq_1/dq_2 representing the relative price of the inputs $-w_2/w_1$. The slope is not affected by proportionate rises in all input prices, only changes in relative prices. Higher isocost lines represent higher input quantities used and higher total costs.

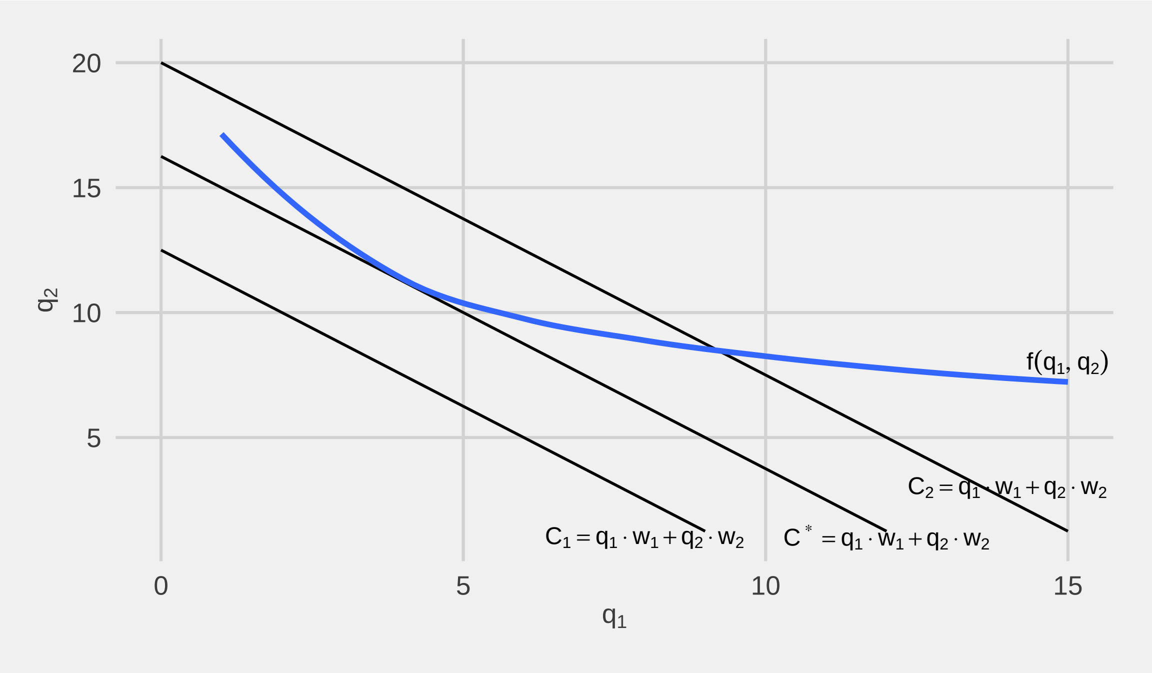

Figure 1.1: Isocost and Isoquant Lines. Isocosts C(q,w) are linear and represent bundles of inputs with identical total costs. Isoquants f(q1,q2) are curved and represent bundles of inputs which produce the same output. The organisation chooses input use at a point of tangency between the two, for example, at q1 = 4, q2 = 12, on isocost line C*. This output choice reaches the highest isoquant for a given cost or the lowest isocost for a given level of production

The curved isoquant lines are convex and represent fixed levels of output y, determined by the production function $y = f(q_1, q_2)$. Isoquant lines also reflect fixed levels of revenue $p_y \cdot y$ as the product of output and its price p_y, assuming the organisation is price-taking with limited market power. The slope of each isoquant is derived by taking the derivative of the production function with respect to each input $dq_1/dq_2 = -(\partial y/\partial q_2)/(\partial y/\partial q_1)$

The standard model of rational behaviour requires the organisation to choose an output level represented by a particular isoquant line. Having done so, the organisation must minimise costs for that output level. To minimise cost, the organisation selects the position on the possible lowest isocost line for the selected level of output. The optimal position is accomplished by choosing a point on the isocost line tangential to the isoquant line, where no possible combinations of inputs yield the required production level at a lower cost. This point is achieved where $w_2/w_1 = (\partial y/\partial q_2)/(\partial y/\partial q_1)$. The slope of the isoquant line (the ratio of marginal products of each input) must equal the slope of the isocost line (the ratio of costs of each input). Isocost lines below this point, for example C_1, are cheaper but do not yield the required level of output. Isocost lines above this, for example C_2, are more expensive. The organisation minimises costs subject to output at the point of tangency between isocost and isoquant lines.

1.2.3 Short-run and Long-run Costs

The example above assumes that the organisation can immediately vary all inputs. However, the organisation may be unable to vary some factors of production in the short-run. For example, constructing a hospital building would take considerable time. In the short-run, the supply (and associated cost) of these factors is fixed. Fixed factors of production can be thought of as a sunk cost that does not vary with output. In the long-run, a hospital organisation can construct new buildings, move to new sites, or otherwise amend the capital stock embodied in the building. Changing the capital stock employed will affect expenditure, and consequently, organisational costs can differ in the short and long-run.

The short-run position is set out in Figure 1.2. If the input q_1 is fixed at $q_1 = 5$, short-run total costs can be expressed as $C_{srun} = FC_{q_1=5} + q_2 \cdot w_2$. Here, FC is a fixed cost associated with the current level of the fixed input, equal to $5w_1$. This fixed cost FC is independent of the level of output chosen. The usable level of q_1 may be less than 5 if the production technology means the marginal product of capital declines to 0 after a certain point. However, the organisation cannot use more of input q_1 than is available in the short-run. With q_1 capped, the only decision for the organisation is to choose the level of input q_2 to obtain the desired level of production on the relevant isoquant line. Mathematically, this decision can be expressed as $min\, q_2 \cdot w_2 \quad s.t.\, f(q_1, q_2) \geq y'$, where $f(q_1, q_2)$ is the production function and y' the desired level of production. This optimisation decision yields the short-run cost function $C^*_{srun}(q_2, y, w_1, w_2)$ which is expressible as a function of the variable input q_2, the required production level y and input prices w_1, w_2.

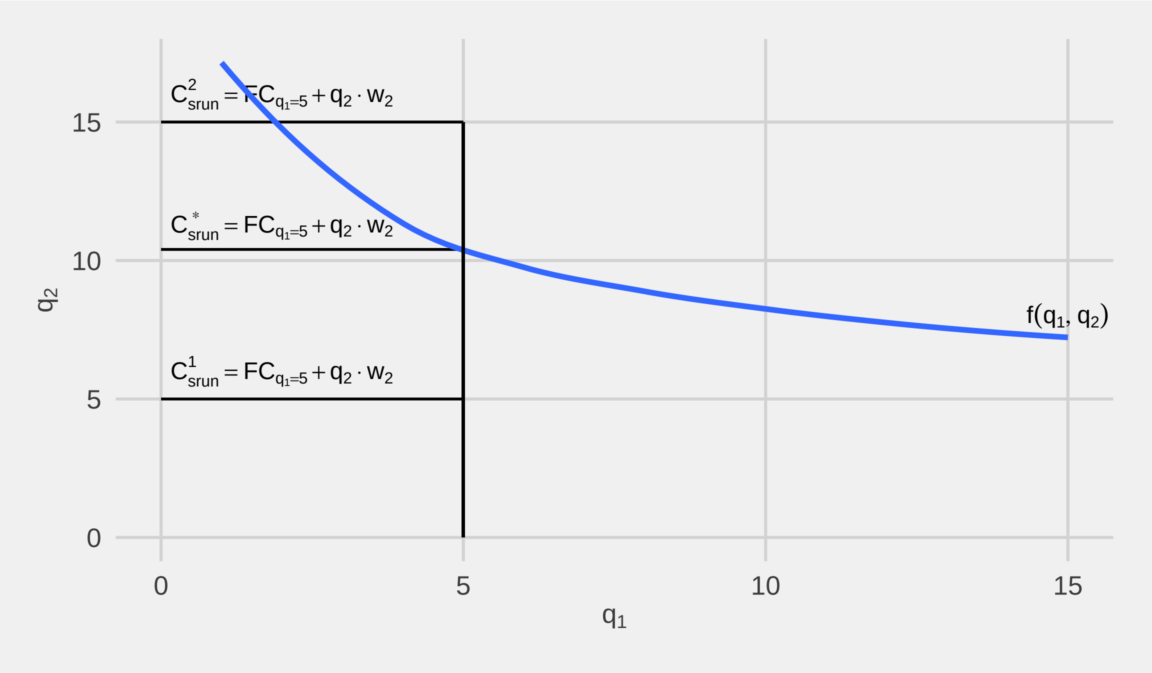

Figure 1.2: Short-run Isocost Lines. Input q1 is fixed at q1 = 5. The organisation chooses q2 to reach the desired isoquant f(q1, q2), achieved with lowest cost at the isocost line C* srun

In the long-run, where all inputs are variable, the long-run cost function reverts to $C(y, \mathbf{w})$, the minimum cost required to produce output y given a production technology and input prices $\mathbf{w}$.

Having established that overall costs can differ in the short and long-run, it is helpful to consider how average costs are likely to change dependent on the timeframe considered.

1.2.4 Average Costs in the Short and Long-run

Average cost curves are assumed to be ‘U’ shaped when plotted against output, starting at a higher level and then declining as y increases, before rising again after a certain point. The rise and fall may be the result of scale effects. For example, low output levels are dominated by large, fixed startup costs, which do not increase with additional output. After a certain point, diminishing returns arise, perhaps due to coordination problems at higher output levels. Average costs start to increase again, creating the ‘U’ shape.

The ‘U’ shape of average cost curves will differ in the short and long-run. In the short-run, the inability to amend fixed factors of production could cause marginal and average cost curves to be higher than in the long-run, where all factors of production are variable. In the long-run, all short-run positions are available to the organisation, so the organisation can produce at an average and marginal cost that is as low as the best short-run curve. Therefore, long-run average costs should be tangential to or below each of the short-run curves (Figure 1.3). The long-run average cost function is an envelope of the short-run average costs.

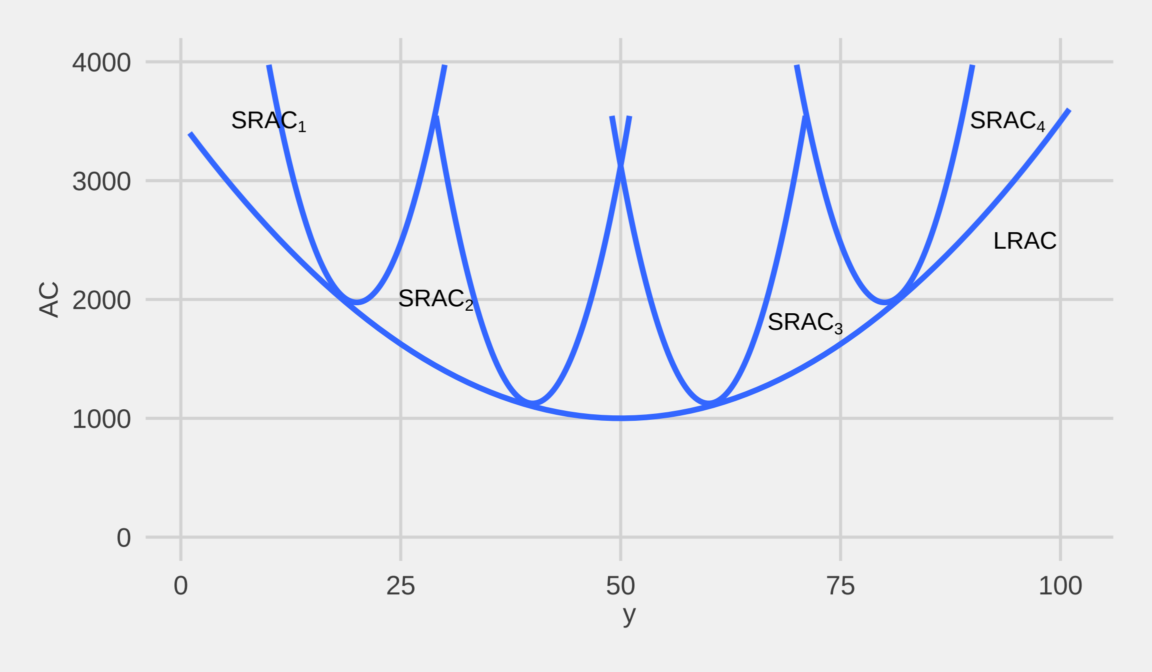

Figure 1.3: Long-run and Short-run Average Cost Curves. Each point on the long-run average cost curve is the lowest point on a short-run curve. In the long-run, the organisation can choose inputs and production technology to do as well as the optimal short-run allocation.

1.2.5 Cost-Output Elasticities

It is useful to develop the concept of how cost changes with output to investigate the relation of scale to cost. The elasticity of cost with respect to output tells us the percentage change in cost associated with a 1% increase in output. Note that this can also be expressed as the ratio of marginal costs $MC = dC(w, y)/dy$ and average costs $AC = C(w, y)/y$ (Gravelle & Rees, 2004, p. 121).

$$\frac{dlnC(w, y)}{dlny} = \frac{dC(w, y)}{dy} \cdot \frac{y}{C(w, y)} = \frac{MC}{AC} \tag{1.4}$$

The cost-output elasticity and the ratio of marginal and average costs can be used to measure the degree of scale economies, as discussed below.

1.2.6 Scale Economies

In the single-product case, the degree of scale economies is typically defined according to the slope of the average cost curve over a particular output range. That is, whether average costs are falling, constant or rising as output increases. Where the average cost curve is falling with increases in output, higher output is associated with lower costs per unit, implying positive scale economies. Where the average cost curve is rising, increases in output are associated with higher costs per unit, defined as negative scale economies. The range or point where average costs are flat or do not change markedly with output is often referred to as the 'minimum efficient scale' or MES. This minimum efficient scale can be equal to or greater than total market demand if average costs do not rise before total market demand is met. This scenario is often termed a 'natural monopoly'.

The degree to which scale economies are present is measured by the Scale Economy Index S, defined by Baumol et al. (1982) as the ratio of average to marginal cost AC/MC. This scale economy index is the reciprocal of the cost-output elasticity equation defined at (1.4). This index tells us whether average costs rise or fall at a given point. To see this, note that the average cost curve will rise where marginal costs are higher than average costs and fall where marginal costs are lower than average costs. The marginal cost curve will intersect the average cost curve at the lowest average cost.

The ratio of marginal and average costs at each output level determines the local behaviour of the average cost curve. Where the marginal cost curve is below the average cost curve, costs decline with output. Each additional unit of output adds a unit cost lower than the cumulative average cost, lowering average costs and causing the AC curve to slope downwards. In this case, the index AC/MC is greater than one, implying positive scale economies. If the marginal cost curve is above the average cost curve, the reverse is true, with each additional unit of production added at a cost greater than the current average. Where this is true, the average cost curve rises with output, and there are negative scale economies as $S < 1$.

Under the standard microeconomic theory, average cost curves are expected to have a 'U' shape with respect to output. Organisations exhibit positive scale economies initially as fixed costs are spread over increasing output before flattening out and then showing diseconomies of scale. Diseconomies of scale may arise where an organisation finds coordinating itself difficult or cannot operate efficiently at higher output levels.

1.2.7 Desirable Properties of Cost Functions

It is helpful to define some properties of the cost function required for consistency with microeconomic theory. Ideally, any estimated cost function should conform to these desirable microeconomic properties. For example, Chambers (1988) sets out six conditions:

1. **Cost should be non-negative** — $c(w, y) > 0$ where $w > 0$ and $y > 0$. This condition can be evaluated by the form of the cost function.
2. **Cost should be non-decreasing in input prices** — $\partial c(w, y)/\partial w_i \geq 0$. This condition can be evaluated by differentiation of the estimated function.
3. **Cost should be continuous and concave in input prices** — $c(w, y)$ is continuous and concave for any w_i. This condition follows from condition 2 and the possibility of input substitution and cost-minimisation. To see this, consider the case of input i. In the event of a rise in price from w_i to w_i^* and no input substitution, total costs would increase by $(w_i^* - w_i).q_i$, where q_i is the quantity of input i employed. Total costs would increase by less than this if the organisation can find alternatives which result in lower costs. However, a cost-minimising organisation would never switch to an input bundle such that the total cost increased by more than $(w_i^* - w_i).q_i$. This cost-minimising assumption ensures weak concavity of $c(w, y)$ with respect to any given w_i. This requirement can be evaluated by calculating the Hessian matrix of partial derivatives with respect to input prices. Where the Hessian matrix is negative semidefinite, the cost function is concave with respect to input prices.
4. **Cost should be non-decreasing in any output** — $\partial c(w, y)/\partial y_i \geq 0$. This condition follows from the assumption of a monotonically increasing production function. Such a production function requires minimum inputs to rise with increases in output, hence raising the minimum cost. This property can be checked by computing the partial derivatives of cost with respect to each input price or the relevant cost-output elasticity, which will have the same sign. Where the result is positive, costs increase with output.
5. **Cost should be linearly homogeneous in input prices** — $c(tw, y) = tc(w, y)$. This condition states that minimum costs rise proportionately when all input prices rise together. As any opportunities for input substitution are conditional on changes in *relative* input prices, there can be no cost-minimising substitutions where all input prices increase by the same amount. This condition could be enforced by restrictions on the estimation method, where econometric software allows. However, it is worth noting that restrictions on coefficients to enforce either this condition or condition #4 may result in distortionary effects on marginal costs and scale economy estimates (Brown et al.,

1979). There is perhaps a tradeoff between theoretical consistency and goodness of fit that may be as much a methodological preference as an ontological question.

6. **No fixed costs in the long-run** — $c(w, 0) = 0$. Where all inputs are variable, cost should equal zero when no output is produced. This condition can be either evaluated by substitution of a zero vector into the output variables and computing costs or tested asymptotically by taking the limits of the cost function as $\lim_{\sum y_i \to 0+}$

These properties can be used to evaluate how well estimated cost equations fit standard microeconomic theory and the internal consistency of their results. Ideally, all these properties would be observed in estimated cost functions. Departures from these properties would indicate that the estimated cost function was logically inconsistent, less likely to behave as expected, and consequently less desirable empirically.

1.3 Multiproduct Organisations

1.3.1 Multiproduct Average Costs

The definition of scale economies relating to the behaviour of average costs is straightforward in the single-product case. However, this simple formulation cannot be used where organisations produce more than one good. Following the work of Baumol et al. (1982), it is noted that scale economy definitions based on average costs become complex when there are multiple outputs. With multiple outputs, the average cost curve becomes a multidimensional surface rather than a line. Defining average cost as $AC = C/\mathbf{y}$ and marginal costs $MC = dC/d\mathbf{y}$ becomes complex where $\mathbf{y}$ now represents a vector of multiple goods.

In the first instance, average costs need an alternate definition in the multiproduct case, as there is no natural way of combining a basket of outputs into the denominator $\mathbf{y}$ to average. One common approach to this problem is to hold the proportion of outputs constant at the current output mix. For example, see Koenker (1977) for an early example applied to road transport. When the proportion of outputs is fixed, percentage increases in this fixed basket of outputs are easily computed. Following this method, the multiproduct case is redefined to a single composite good. This redefinition allows average costs to be computed with reference to the basket, usually defined as current output, as a ray average cost.

Ray Average Costs in this context were first defined by Baumol et al. (1982). Where the bundle of goods has been defined as y', a ray average cost is defined as $C(ty', w)/t$, with t as the number of units of the composite good in the bundle. This expression measures the slope of the plane (or ray) from the origin to the defined bundle of goods multiplied by whatever level of output scalar t is used. Consequently, average costs can be defined in the multiproduct case and elasticities and scale economies can be computed. As the proportion of inputs along each ray is unchanged, the bundle of goods remains the same across different outputs, avoiding the difficulties of combining different outputs.

A disadvantage of this approach is that comparisons at the industry level are problematic, as each organisation's output mix will differ. A finding that ray average costs for a particular organisation continually fall does not necessarily mean that the industry is a natural monopoly, as another organisation with a different output mix may experience a different cost curve. However, we can evaluate the general behaviour of each organisation's average costs relative to current outputs, and infer sector-level behaviour from how each of these individual organisations' cost curves behaves locally.

1.3.2 Multiproduct Marginal Costs

The definition of marginal costs also becomes more complex in the multiproduct case. In the single good case, marginal costs are the incremental costs of one more unit of the only good. In the multiproduct case, the marginal costs of one good can be affected by the production level of other goods. The analogous definition for marginal costs becomes the partial derivative of cost with respect to the output concerned $\partial C(w, y)/\partial y_i$. Whilst this retains the idea of measuring marginal changes with respect to output, it makes multiproduct marginal costs conditional on the cross-products of y_i and other outputs $y_{n \neq i}$. Marginal costs thus defined depend on the current level of other outputs, so they are no longer unique to particular levels of y_i.

1.3.3 Multiproduct Scale Economies

Adopting these definitions of (ray) average and marginal costs allows the construction of a scale economy definition analogous to the single-product case. Scale economies are defined by whether average costs rise or fall over a defined output level. However, because our definition of ray average cost includes multiple outputs, it does not permit either a simple derivative test for whether overall average costs are rising or falling, or a natural scale over which to evaluate this. Scale economies must therefore be calculated based on elasticities and the ratio of average and marginal costs. In the multiproduct case, the scale economy measure $S = AC/MC$ becomes the inverse of the sum of individual cost-output elasticities, as follows:

$$\begin{aligned} S &= \frac{AC}{MC} \\ &= \frac{\left(\frac{C(w,y)}{\sum_i y_i}\right)}{\sum_i \frac{\partial C(w,y)}{\partial y_i}} \\ &= \frac{C(w,y)}{\sum_i y_i \cdot \frac{\partial C(w,y)}{\partial y_i}} \\ &= \frac{1}{\sum_i \epsilon_{C,y_i}} \end{aligned} \tag{1.5}$$

Where $\epsilon_{C,y_i} = \partial \, ln \, C(w,y)/\partial \, ln \, y_i$ is the cost elasticity of the ith output. This equation gives us a definition of scale economies for the multiproduct case, avoiding the issues of comparing qualitatively different goods.

1.3.4 Jointness and Economies of Scope

The multiproduct case also raises the complication of joint production, which would not be applicable in the single good case. Joint production occurs where the cost of simultaneously producing two or more outputs is less than the cost of producing them separately. Mathematically, joint production for an organisation producing two goods is present where the following inequality holds:

$$C(y_1, y_2, w) < C(y_1, 0, w) + C(0, y_2, w) \tag{1.6}$$

Instances where this inequality applies are defined as ‘sub-additive’. If the inequality is true over all values of y up to the market demand, then the cost function is globally sub-additive and a natural monopoly. Costs are lower with production concentrated than in the case of separate production.

If jointness is absent and the cost of independent production is equal to joint production, the production costs of each good are independent of the output levels of other goods. Here, production activities are functionally independent, and the organisation could be separated into individual organisations or sites without affecting the cost of production.

Jointness in production is closely related to economies of scope. Economies of scope may be present where the production of different outputs can use shared resources or gain benefits from colocation or co-production. For example, the marginal costs of one surgical specialty may be affected by investment in theatres and postoperative beds for another surgical specialty. Diagnostic imaging equipment bought for elective daytime work may be used for emergency out-of-hours cases at a lower marginal rate than purchasing another set of equipment. Each of these examples corresponds to cases where inequality (1.6) applies. This definition is used in Chapter 5 as the test for the presence of economies of scale.

1.4 Measurement and Definition of Hospital Cost Concepts

1.4.1 Defining a Hospital

It is first useful to define what a 'hospital' is, firstly as a means of defining an appropriate study population, and secondly to define the 'decision making unit' (DMU) or the level at which scale economies are measured.

For the first criterion, I define a hospital as a provider of inpatient care across multiple clinical specialties, including both emergency and planned care. In England, these organisations are generally referred to as 'acute providers'. In addition to inpatient care, these organisations offer healthcare in outpatient settings, diagnostic imaging, emergency department care and other services. Specialist organisations offering only planned inpatient care in a restricted range of specialties are excluded - see Appendix A for details. Also excluded are organisations that only provide mental health services, private sector organisations that do not offer emergency care, and community care organisations that only offer services such as health visiting and district nursing. The organisations that fit this definition give a study population offering comparable healthcare in a broad range of clinical specialties.

To see why the second criterion is important, note that acute provider organisations may run more than one hospital site. In this respect, it is useful to define whether we are considering the site or the organisation level when estimating economies of scale and scope. In other countries, a hospital organisation could operate as a hospital chain or network, with several self-contained sites and significant decision making delegated to each site. Where organisations span multiple sites, it might be argued that the site should be considered as the DMU. However, it is more natural to consider the organisation level as the unit of production where decisions on output levels are centralised at the organisation level, or if each site is not self-contained and specialises in different types of healthcare.

In the English case, many acute organisations run multiple sites. However, decisions on capital investment and healthcare output are taken at the organisation level. In addition, where an organisation has multiple sites, it may also centralise particular services to one of these sites. This centralisation minimises the costs of staff moving between sites and the costs of duplicating smaller departments on each site. A multi-site configuration means that the full range of hospital services is provided at the organisation rather than the site level.

Finally, using a site-based definition of the DMU also means difficulties allocating output in rarer cases where departments span different units and patients may receive treatment in both. For example, a patient may attend one site for surgery and another site subsequently for outpatient follow-ups. For these reasons, in this thesis, the organisation is the decision making unit and costs are examined at this level of aggregation.

1.4.2 Healthcare or Health?

A fundamental issue of output measurement is whether to measure healthcare provided or health improvements realised. A more comprehensive discussion of each side is in Butler (1995) pp 48-55. From the patient's point of view, the desired end state is an improvement in their health. A health improvement may need to be set off against utility losses from side effects or other negative externalities such as travel. However, assuming these utility losses are minimal, the health improvement is the individual's main 'good'. Consequently, there is a school of thought which argues that output should be measured in terms of the increase in health obtained from each hospital visit.

Practically, measuring such a health improvement is difficult. Widespread recording of health status measures such as EQ-5D (Euroqol, 2022) for each treatment does not occur. Government institutions may model quality-of-life improvements as part of cost-effectiveness analyses for new treatments. However, there is no database or systematic retrospective measurement for existing treatments. Measurement is even more difficult for complex pathways involving multiple treatments, visits, and interactions. Average health improvements calculated at a certain point may change with time as treatments are refined and life expectancies change. Consequently, even if it is desirable to define hospital output in terms of health, measurement and data deficiencies make this approach unfeasible.

The alternate view of measuring healthcare output could also be regarded as more theoretically correct. The output of other industries is not usually measured by their effects on the individual. Even if measurement were possible, there are situations where output as health improvement results in apparently perverse implications. For example, Butler (1995) notes that taking health state improvements as output means that palliative care results in no output. However, the issue is even more general than this. Palliative care could be said to offer a health improvement over the alternative of no palliative care if pain and discomfort are included in the definition of health. However, health improvements (over time) need to be evaluated against a hypothetical alternative. For example, a patient receiving treatment for eye cataracts is expecting an improvement in vision compared to the day before surgery. However, this health improvement is valued not only for its effect at the time of surgery. The health status improvement is relative to the expectation that vision could deteriorate further without treatment. Such counterfactuals are difficult to compare on any significant scale as they may be contingent on life expectancies or other comorbidities. We would need to use an average treatment effect per healthcare procedure to measure this, which takes us closer to measuring healthcare in any case. It is theoretically arguable whether hospital outputs should be measured as healthcare or health improvements. However, our need to measure health outputs on a large scale means that measuring healthcare as the output is the only practical empirical method.

1.4.3 Aggregating Healthcare Outputs

Having defined a hospital and the form of outputs, it is necessary to decide how to aggregate outputs to a manageable level. Cost functions with few parameters have the advantage of parsimony and are easier to estimate, but have the drawback that the aggregated outputs are harder to relate to real-world policy interpretations and do not capture nuances of differing patient casemix (Breyer, 1987). Some degree of aggregation of differing healthcare goods is necessary, as individual patients can have different treatments depending on their circumstances, comorbidities, and personal preferences.

Early studies counted the number of cases or the number of bed days (Lave & Lave, 1970). Counting bed days attempts to account for a degree of complexity in casemix. More resource-intensive healthcare would be associated with a longer stay in hospital. However, this association is not perfect, and counting outputs by bed days only applies to inpatient care. Counting bed days cannot account for complexity in other settings. Bed days may also be a poor proxy for complexity. A long-staying patient may represent a greater intensity of healthcare usage than a short-staying one, dependent on the qualitative details of the treatment and required recovery time. Using bed days as output also makes no allowance for inpatient stays that are longer than necessary due to organisational inefficiency rather than clinical requirements. Where hospitals are inefficient, length of stay may be longer (Siciliani et al., 2013). Even with the same treatment and identical hospital characteristics, the length of hospital stay can also vary depending on the patient's general health or the availability of subsequent community or social care. Where stay lengths are long, the marginal difference in additional case cost may be minimal if the additional days only provide nursing and 'hotel' costs with limited additional treatment. Using bed days as outputs could mean that a hospital with longer average stays records higher output and lower average costs, despite providing the same amount of treatment except for accommodation and nursing. Using bed days as a method of accounting for complexity or as the basis for measuring output may be better than crude counts of cases, but it is still a poor means of categorising and measuring outputs. There are alternative methods of aggregating and classifying similar types of healthcare that can be used to take account of heterogeneity in healthcare provided, adjusting for qualitative differences in care provided and differences in casemix.

1.4.4 Casemix and DRG/HRGs

Butler (1995) discusses two main casemix classification schemes, The International Classification of Diseases (ICD) and Diagnosis Related Groups (DRG). The ICD taxonomy organises and provides classifications based on diagnosis or the particular health need to be met. For example, ICD-10 (the edition used in the UK at the time of writing) includes 22 chapters (WHO, 2022). ICD groupings, based on diagnoses, provide a means of classification based on health needs but not on the healthcare provided.

DRG groups classify according to diagnosis and also treatment options. DRG classifications are (in the US Medicare System) subdivided into Major Diagnostic Categories based on diagnoses, and subdivided further based on the treatment offered for that diagnostic group. Within this category are DRGs for Percutaneous Cardiovascular Procedures involving artery stenting, with differing DRG codes depending on whether the stent is drug-eluding, whether the patient has comorbidities, and the degree of seriousness of these, and the number of stents used in the procedure. Relatively detailed information is encoded within these DRGs, including the patient's general health and the healthcare to be provided.

In the UK, similar groupings to DRGs are used and named 'Healthcare Resource Groups' or HRGs. HRG codes were developed with a similar intent to DRGs, as a means of administering government payments to hospitals and incentivising cost control, rather than funding via cost reimbursement or block grant. As in the US, HRGs are determined by rules based on the appropriate diagnosis and procedures carried out as part of care (NHS Digital, 2022a). UK HRGs also take into account the degree of comorbidities present where applicable. Reimbursement is based on the care setting and HRG used, so an HRG may attract a different payment depending on whether an inpatient stay is required. HRGs are used in subsequent chapters to weight and account for differences in casemix and output heterogeneity, as they can capture differences in treatment provided, as well as differences in healthcare needs.

The number of HRGs precludes using individual codes as output categories, but it is possible to use historic costs or reimbursement amounts as a means of weighting when aggregating to a higher level. Outputs can thus be aggregated to a level where estimation becomes feasible and models are sufficiently parsimonious. Outputs are aggregated to the care setting level in Chapters 2-4 and the specialty level in Chapter 5, dependent on the factor of interest, which is scale economies in the former and scope economies in the latter.

1.4.5 Incentives and Coding Quality

Using HRGs as a means of aggregation offers a more granular method of dealing with heterogeneity, but HRGs are not without drawbacks. The HRG classification and reimbursement system is used primarily for cost control but also to incentivise certain treatment types. Where this happens, an HRG-based weighting that uses reimbursement values could be distorted. For example, certain day case surgery is incentivised by being funded at the same amount as patients who stay in the hospital overnight, despite the former arguably receiving more healthcare. This incentivised pricing can lower costs (Gaughan et al., 2019), but also means that if reimbursement values are used for weighting, healthcare outputs could be distorted.

HRG allocation can be affected by clinical coding practices. Hospitals that 'undercode' or do not accurately record diagnoses, comorbidities, or treatments record lower value HRGs than hospitals that better capture this information. Coding quality may also be lower at smaller hospitals, which may be a particular issue where size is used as an explanatory variable. Smaller hospitals with limited resources may lack funds to hire sufficient clinical coding professionals to record codes. They may also be unable to purchase health records software to help automate the process. Lacking resources could be a self-reinforcing situation. Where poor quality coding is present, hospitals may not be reimbursed fully according to the healthcare supplied. As well as undercoding hospitals not recording aspects of care, there are incentives for 'HRG Creep', coding for higher value HRGs where rules are ambiguous. Evidence for this occurring in practice is limited, though instances would be difficult to detect (Rogers et al., 2005).

1.4.6 Non-Healthcare Outputs

As well as healthcare outputs, hospitals can also produce goods in the form of undergraduate medical and nursing training. Hospitals that provide teaching have higher costs (Farsi & Filippini, 2008). Many studies attempt to measure the impact of teaching on hospital costs, that is, they estimate the premium on overall cost arising from teaching hospital status (Butler, 1995, pp. p217–248). I do not attempt to add to these studies, as the cost dataset used attempts to exclude teaching costs. A dummy variable for teaching costs is used to account for this exclusion being only partially successful. Teaching activities are likely not entirely separable from treatment activities as instances of joint production. For example, ward rounds with medical trainees and similar activities involve healthcare and teaching. Teaching hospitals are also likely to be larger hospitals treating a wider variety of more resource-intensive healthcare needs, so they have a confounding issue associated with casemix. The alternative, constructing a dataset inclusive of teaching costs and accounting for the joint nature of production, requires using financial statement values which include other miscellaneous income and expense, accounting adjustments and other potential sources of difference. Any attempt to model teaching outputs alongside those of healthcare has to deal with the challenging issue of devising a conceptual unit to measure teaching, other than crude indicators like the number of students at the hospital.

Some hospitals also undertake significant research activities, which can also increase costs (Linna & Häkkinen, 2006). Research activities also have a similar issue to teaching activities in that activities must include an element of joint production. A patient taking part in a trial or other treatment research must receive some treatment, either a new intervention or standard care. As with teaching activities, higher research activity is likely to occur at larger hospitals with more complex casemix and other qualitative differences which facilitate research.

In subsequent chapters, data is taken from the English National Cost Collection Exercise (NHS England, 2022b), formerly Reference Costs. This dataset attempts to exclude teaching and research costs via a process based on apportionment and staff timesheets. Given the presence of joint production, this is likely to cause distortions, but there is no other cost dataset (such as financial statements) that does not have similar issues separating the cost of non-healthcare goods from healthcare production.

1.4.7 Defining Size and Measuring Capital

To investigate the relationship between cost and size, a definition of hospital size is needed. Most studies avoid measuring scale economies directly against output to avoid the version of the “regression fallacy” identified by Friedman (1955). Friedman notes that organisations may have ‘random’ fluctuations in output in a cross-sectional dataset despite constant fixed costs. The element of average cost attributable to fixed factors of production would vary inversely with output, potentially biasing attempts to examine estimates of 'optimum size'. Consequently, most historic studies have used bed numbers to measure size (Giancotti et al., 2017).

However, bed numbers are imperfect measures of capacity, especially over time. There has been a longer-term trend for non-admitted healthcare to increase as a share of hospital output and an associated reduction in the length of hospital stay and associated bed numbers (McKee, 2004). In addition, using beds as a proxy could introduce other biases. For example, hospitals with many beds relative to output may be inefficient rather than inherently 'large'. Inefficiency amongst hospitals with higher bed numbers could cause a potential bias in the opposing direction, understating scale economies. Despite this, there are few alternative measures of size available other than output measures, so bed numbers continue to be widely used in the literature.

Some studies which employ an envelope method to derive long-run cost functions use proxy variables for capital employed (Kristensen et al., 2012; Preyra & Pink, 2006). These studies have typically also used bed numbers as a proxy for capital. Following a similar methodology in the case of England would also require a proxy value for economic capital, as there are no established capital markets in the UK for hospitals. Data available offers two possible measures to use as a proxy for a capital or 'size' measure. Either bed numbers or financial statement fixed asset values can be used as a proxy for capital or size. Data sources for capital measures are discussed below.

1.4.7.1 Bed numbers

Hospital bed numbers are taken from publicly available data submitted by English NHS hospitals. Each NHS hospital must complete a periodic audit of beds and occupancy levels and submit these to NHS England, with data published on NHS England's statistical work areas website (NHS England, 2019). 'Beds' are defined according to the NHS data dictionary definition (NHS England, 2022a), which sets out the difference between beds, trolleys, and couches. Bed data is separated into 'Day only' beds and Overnight beds. Day beds are used for patients whose care is planned so they can be admitted and discharged on the same day, with the bed being unavailable overnight. Overnight beds are used for overnight stays, where a patient is admitted and remains in the hospital for at least one night before discharge. Both bed types are combined into a total bed number for analysis. Quarterly bed numbers are extracted for each hospital and then averaged over each year to get an annual average bed number for each hospital-year data point.

How the different types of beds are counted is open to debate. Arguably, day beds could be weighted to a lower extent than overnight beds, as their capacity is limited by the hours they are available. However, day-case activity is already weighted according to the HRG reference cost values. Hospitals with a large number of day beds relative to overnight beds will therefore have lower weighted activity relative to hospitals with fewer day beds relative to overnight beds. I do not weight beds according to type for three reasons. Firstly, any weighting of day beds relative to overnight beds would be arbitrary. One option would be to count day beds as 50% of the value of an overnight bed, approximating the hours that they would be 'open'. However, the hours that the day beds are open varies according to individual hospitals and even within hospitals. Even for overnight beds, most healthcare activity would happen during the day. There is no obvious way to weight day beds relative to overnight beds. Secondly, the capital embodied in the bed is the same irrespective of usage. A day case ward would still require similar infrastructure and capital investment. The main cost difference between the two types of beds is the labour (predominantly nursing) used in conjunction with the bed. Day only beds will not require nursing labour during the hours the beds are not in use. Finally, hospitals can convert beds between day and overnight if desired, subject to need and available staffing, making the distinction partly arbitrary in practice.

Using bed numbers as a proxy for capital has some drawbacks. Notably, bed number requirements are set according to the amount of inpatient care required. If a hospital has a higher than average requirement for inpatient care, its bed numbers will be higher than those of an otherwise comparable hospital which provides more ambulatory care in outpatient clinics. Nor are comparable bed numbers indicative of the same level of capital intensity, even amongst inpatient care. For example, critical care beds may be associated with higher capital levels than 'standard' beds. This additional capital is due to critical care beds requiring more investment in ancillary equipment than standard ward beds. Hospitals with higher numbers of critical care beds are also likely to carry out higher numbers of surgical episodes, as major surgery may require a period of postoperative critical care. More critical care beds or beds for surgery will require additional capital embodied in surgical theatres. Using total beds as a proxy for capital could therefore be misleading where there are differences in the *type* of bed used.

Bed number requirements are additionally affected by clinical reasons which are beyond the scope of the model presented here. Hospitals that treat a larger number of non-elective patients (emergency care admissions and births) may need to keep a higher 'buffer' of beds which may often be vacant but are available in the event of a large number of simultaneous emergency cases requiring care. Such hospitals may need a higher level of capital employed for a given activity level to deal with brief periods of high demand. Hospitals with smaller numbers of non-elective patients may not need to flex their capacity and may consequently have lower numbers of average beds, even if they treat comparable numbers of patients. Separating non-elective and inpatient care in the model should at least partially account for this. However, hospitals with more volatile demands for non-elective care could still have higher bed numbers and capital requirements than hospitals with more predictable or lower non-elective care demand.

Bed requirements can also be influenced by the capacity of community and social care organisations. Patients who are medically fit to be discharged may have delayed discharges if there is no social care resource to support them (Holmås et al., 2013). Consequently, hospitals may require higher bed numbers depending on their region's community and social care capacity.

1.4.7.2 Financial Statement Asset Values

The alternate measure of size or capital would be to use capital values from published financial accounts. Accounting measures of fixed assets can be obtained annually from published financial statements. Financial statements include accounting measures of capital as part of their statements of financial position, freely available via the Trust Accounts Consolidation ('TAC') datasets published annually by NHS England (NHS England, 2022c). Financial statements include sections for long-run assets, typically land and buildings, and short-run assets, such as debtors, stock, and money in bank accounts. Many of these short-run assets would not typically be considered assets in the economic definition of capital, so they were excluded. Long-run assets in hospital financial statements are typically dominated by land and buildings, whose valuation is much larger than short-run assets. Consequently, capital values from the long-run assets section of the accounts are closest to the economic definition of capital. Using long-run assets also ensures that the majority of assets by value are included in the measure.

Financial statement valuations of capital are inherently problematic for several reasons. Accounting definitions of capital may not correspond to economic conceptions and valuations of capital. Accounting valuations can be particularly sensitive to cases where assets are rented and not owned, as financial statements only count owned rather than rented assets. In some situations, accounting definitions of 'ownership' can be a grey area, for example, assets purchased through the Private Finance Initiative ('PFI') scheme. The PFI scheme enabled hospitals to make large-scale investments funded by commercial finance rather than public investment. Often these involved hospitals not taking ownership of an asset but paying a finance and service charge each year. As the asset would not be under the hospital's ownership, it would not appear in the assets section of the hospital financial statements. Comparable assets could be included for one hospital but excluded from another, depending on financing.

Even where ownership and financing are comparable, the financial valuation of assets can also be a source of variation. Capital valuations for land and buildings cannot use market valuations as no developed private markets exist for comparable assets (hospital buildings) in the UK. Financial statements consequently include assets at a value arrived at by a specialist valuer. These valuations are carried out by specialist valuers who value property capital according to the cost of constructing an equivalent asset generating equivalent utility. That is, they value according to the cost of constructing a comparable hospital in a comparable location. Financial capital values reflect a hypothetical alternative rather than a market value. Valuations can also be affected by professional opinions on what is an 'equivalent' utility or location. Taking all these factors together introduces an element of variation according to the judgements of professional valuers. How well this reflects 'true' capital values has significant uncertainty. Using financial statement values of capital also has a practical drawback when comparing results across time or country. There is no obvious deflator for hospital valuations across time, and cross-country comparisons introduce additional uncertainty from exchange rate comparisons.

Both bed numbers and financial statement values have significant drawbacks as proxy measures of capital. In this thesis, bed numbers are used as the preferred measure of capital due to the inherent variability of financial statement capital accounting and the artificial nature of hospital property valuation in the absence of a developed market for hospital buildings.

1.5 Estimation

1.5.1 Parametric or Non-Parametric Methods

Studies considered thus far use parametric methods to estimate a cost function. There is an extensive literature analysing hospital production and efficiency using non-parametric methods, chiefly using Data Envelopment Analysis (DEA). Surveys of this literature include Hollingsworth (2003); Hollingsworth (2008); Kohl et al. (2019) and O'Neill et al. (2008). Few studies use both parametric and non-parametric methods on the same data. Siciliani (2006) found comparable efficiency results using constant returns to scale DEA models and Cobb-Douglas parametric models, and more divergence between methods when using more flexible specifications such as DEA variable returns to scale and parametric Translog specifications.

Most of this efficiency literature uses DEA to measure the technical efficiency of hospitals and the relationship of outputs and inputs via the production function rather than scale economies *per se*. However, DEA models can estimate returns to scale, which can be easily related to scale economies. Any production function which exhibits positive returns to scale implies positive scale economies in the cost function, even where the production function is not homothetic (Bell, 1988). However, the reverse is not true - positive scale economies in the cost function, defined by cost-output elasticities, can exist where the production function does not exhibit positive returns to scale. If the production function is homothetic, the relative share of inputs does not change with changes in output, and economies of scale and returns to scale coincide at cost-minimising points. To use the absence of returns to scale as evidence of the absence of scale economies requires assuming a homothetic production function, which may not be justifiable. In older literature, a terminology distinction is made between economies of size (scale economies in the cost function) and economies of scale (returns to scale in the production function), though literature can also use the terms interchangeably. See Chambers (1988) pp. 68-77 for a fuller discussion of terminology, the relationship between returns to scale and scale economies, and their relationship.

DEA approaches estimate an efficient frontier from observed data points using a linear programming approach. This approach derives the frontier of maximum output associated with combinations of inputs, creating an envelope of production possibilities. Efficiency is calculated by how close each DMU is to the frontier as a ratio between 0 and 1. Whilst the basic model assumes constant returns to scale, variable returns to scale can be built into the model by adding a convexity term.

Parametric approaches model the variable of interest (either maximum output y_i, or cost C_i) as a function of x_i (inputs in the case of production functions, outputs and prices in the case of cost functions). Efficiency-based Stochastic Frontier Analysis ('SFA') models the error term with two components. v_i is assumed to be normally distributed with mean 0, and u_i, which measures the distance to the frontier as inefficiency. β are coefficients in the estimation.

$$C_i = f(x_i, \beta) + v_i + u_i \tag{1.7}$$

Much of the DEA and SFA literature is focused on the average efficiency of individual DMUs and the consequent opportunity for efficiency gains. There are few UK-based studies, with most studies looking at US data. To the best of my knowledge, none of the UK-based DEA studies surveyed in Hollingsworth's two review articles report any scale effects at the level of individual hospitals. However, Tsai & Mar Molinero (2002) use a DEA analysis at the medical specialty level, finding an optimal size for surgery but constant returns to scale for maternity services.

The following essays use parametric techniques to estimate a cost function rather than employ DEA frontier approaches. DEA approaches lack the theoretic underpinnings in parametric techniques, simply estimating a frontier which best fits the data. Whilst this flexibility is an advantage in some cases, it removes the ability to apply constraints which may be economically desirable. For example, we may wish to use a particular functional form for our cost function or easily compute comparative statics that may be of interest.

The lack of an error term in DEA also means that DEA approaches can be sensitive to outliers in the data. This issue may cause problems where there are many inputs and outputs, and individual data points may be far away from other observations on the frontier. DEA error terms can be imputed retrospectively through bootstrapping (Simar & Wilson, 1998), though using these methods to fully account for measurement errors is complex.

Finally, though DEA techniques can be used to model returns to scale, their primary use has been to quantify efficiency rather than the effects of scale on cost. It is the latter question in which I am primarily interested. Whilst it is possible to use non-parametric techniques to investigate this, these techniques are rarely used. For the reasons above, parametric techniques are preferred to investigate this question.

1.5.2 Short-run or Long-run Cost Function

Having decided on parametric methods, a subsequent issue is how to model the cost function. The choice is whether to estimate a short or long-run function. Estimating a long-run cost function from direct observations of costs and outputs can be problematic. By doing so, it is assumed that observed cost and output values are cost-minimising for that level of output. However, observed points may not be cost-minimising, for example, due to X-inefficiency (Leibenstein, 1966) or decisions made for reasons other than cost-minimisation. This observation has been the basis for other researchers to critique attempts to estimate long-run functions directly from observed outputs and costs because direct long-term estimation assumes that the observed output levels are short-run optimal points (Aletras, 1999). This assumption is important as if it does not hold, estimates of average cost at each output point could be overstated, reducing the apparent degree of scale economies and adding uncertainty to the model.

Some studies since Aletras (Kristensen et al., 2010; Preyra & Pink, 2006) deal with this issue by first estimating a short-run function where the capital stock is fixed. Subsequently, these researchers generate a long-run function using the envelope of the short-run functions, allowing the capital stock to vary in the long-run. Preyra and Pink use this approach to find no significant long-run limit to scale economies. Kristensen, Olsen, Kilsmark, Lauridsen, and Pedersen reach similar conclusions about long-run cost functions using Danish hospital cost and output data. Estimation results may differ depending on whether a long-run cost function is directly estimated (assuming existing output levels are optimal) or derived via the short-run or variable cost function.

1.5.3 Specific Functional Forms

The final estimation question to be considered is the form of the cost function. The choice of form conditions the behaviour of the cost function. Consequently, it is important to choose a form that both accurately measures the phenomenon of interest and also retains desirable theoretical properties.

Early studies attempting to model a cost function for hospital output tended to use ‘ad-hoc’ specifications. Studies using this ad-hoc or ‘behavioural’ (Evans, 1971) approach included variables for any factor thought to significantly affect cost. Specifications often included variables for bed utilisation, variables to adjust for the heterogeneity of facilities, and other factors that could plausibly affect costs (Hefty, 1969; Mann & Yett, 1968).

After these early studies, most attempts to estimate cost functions used a specification based on economic theory, expressing cost as a function of output and input prices only. Many functional forms have been used to represent cost functions, each with particular characteristics which affect how derivatives such as elasticities behave. A summary of the main forms of cost functions is available in Heathfield (1987). The choice of functional form can therefore affect what can be modelled in the estimation and which properties our estimated cost function has.

Not all functional forms are appropriate for estimating variation in scale economies. The function form needs to allow for the variation of scale economies over outputs. Within the literature, there is a debate around the appropriateness of different functional forms, though few studies test alternate specifications using the same data. Different functional forms have advantages and disadvantages regarding consistency with desirable properties derived from economic theory and the ease or accuracy of their econometric estimation. Vitaliano (1987) finds that these differences can affect estimations of scale economies. The choice of functional form depends on the properties desired in the study, such as homogeneity or variable returns to scale, but we must be mindful that results may change dependent on the functional form used.

1.5.3.1 Cobb Douglas

The Cobb-Douglas form expresses cost in the following form for n outputs and m inputs:

$$C(y,w) = A\, y_1^{\alpha_1} y_2^{\alpha_2} ... y_n^{\alpha_n} w_1^{\beta_1} w_2^{\beta_2} ... w_m^{\beta_m} \tag{1.8}$$

This expression can be linearised to:

$$ln\, C = ln\, A + \sum_n \alpha_n\, ln\, y_n + \sum_m \alpha_m\, ln\, w_m \tag{1.9}$$

More commonly used in production function studies (Sielska & Nojszewska, 2022), using a Cobb-Douglas form has several drawbacks in cost studies. It implies no fixed costs in the non-linear form, as cases where one $y_i = 0$ yields a zero total cost. In the linearised version, any $y_i = 0$ is undefined due to logs. For the function to be linearly homogeneous in input prices requires constraints such that $\sum_m \alpha_m = 1$. The Cobb-Douglas form is homothetic in that it can be written $C(w,y) = A\, c_1(w)\, c_2(y)$, where $c_1 = y_1^{\alpha_1} y_2^{\alpha_2} ... y_n^{\alpha_n}$ and $c_2 = w_1^{\beta_1} w_2^{\beta_2} ... w_m^{\beta_m}$. However, the Cobb-Douglas form does not allow scale economies to vary with output, as returns to scale are a function of the (fixed) coefficients on the output variables $\sum_n \alpha_n$ (Heathfield, 1987, p. 81). Scale economies can be positive, constant or negative according to whether this expression is above, equal to or below 1. However, these scale economies will not vary with output and cannot be used to evaluate the effect of scale on average costs.

1.5.3.2 Flexible Forms

To investigate the relationship between scale economies and size, a functional form that allows varying scale economies is necessary. It is also useful for investigating scope for jointness to be tested after estimating a cost function (M. Fuss et al., 1978). Since Diewert (1971), cost studies have used ‘flexible functional forms’, which can be made to “approximate any function at a point by an appropriate selection of values for the parameters” (Butler, 1995, p. 34).

1.5.3.3 Translog

A widespread form used in cost studies is the transcendental logarithmic or translog form, as outlaid by Caves et al. (1980). This cost function can be written as:

$$\begin{aligned}\ln\, C(y,w) =&\alpha_0 + \sum_i^m \alpha_i \ln\, y_i + \sum_k^n \beta_k \ln\, w_k + \frac{1}{2}\sum_i^m\sum_j^m \delta_{ij} \ln\, y_i \ln\, y_j \\ &+ \frac{1}{2}\sum_k^n\sum_l^n \gamma_{kl} ln\, w_k ln\, w_l + \sum_i^m\sum_k^n \rho_{ik} \ln\, y_i ln\, w_k\end{aligned} \tag{1.10}$$

Where y_i is the amount of the ith output produced (of m total outputs), w_i is the ith input price (of n inputs), and α_i, β_k, δ_{ij}, $\gamma_{k,l}$, and $\rho_{i,k}$ represent coefficients estimated in the model.

The translog has been a popular functional form used when estimating cost equations. Advantages include the ability to model jointness in production and varying scale economies. It is also possible to impose linear homogeneity in input prices via suitable restrictions on the β, γ and ρ coefficients, where these obey the restrictions below:

$$\begin{aligned} &\sum_k^n \beta_k = 1 \\ \forall k = 1,...,n \ :&\sum_l^n \gamma_{kl} = 0 \\ \forall i = 1,...,m \ :&\sum_k^n \rho_{ik} = 0 \end{aligned} \tag{1.11}$$

The main drawback in using the translog, which is true for all flexible forms, is the number of parameters to be estimated. The interaction terms mean that in the multiproduct case, the translog requires $((m(m+1)/2) + (n(n+1)/2) + mn$ coefficients to be estimated, where linear homogeneity of prices is enforced (Caves et al., 1980). The log terms also mean that cost is undefined where certain outputs are zero.

1.5.3.4 Quadratic

The quadratic form is very similar, essentially being a translog form without log transformation:

$$\begin{aligned} C(y,w) =& \alpha_0 + \sum_i^m \alpha_i y_i + \sum_k^n \beta_k w_k + \frac{1}{2}\sum_i^m \sum_j^m \delta_{ij} y_i y_j \\ &+ \frac{1}{2}\sum_k^n \sum_l^n \gamma_{kl} w_k w_l + \sum_i^m \sum_k^n \rho_{ik} y_i w_k \end{aligned} \tag{1.12}$$

Though similar, the lack of log transformation gives two key differences. Positively, using a quadratic cost form means that it is possible to model scenarios where a particular output $y_i = 0$, as this is no longer undefined without the log terms. The drawback is that it is no longer possible to enforce linear homogeneity of degree 1 in input prices through restrictions on coefficients. For the translog, this could be achieved because the log transformation meant that coefficients expressed the effects of percentage changes in input prices. Using the quadratic, this is no longer true, so there is no a priori way of imposing restrictions to ensure homogeneity in input prices.

1.6 Summary

This chapter summarises the key complexities, issues and decisions relating to the evaluation of hospital cost functions. The multiproduct nature of hospital healthcare means a more complex means of modelling costs is needed, compared with the single-product case. The joint production of healthcare goods means costs cannot be apportioned to individual products, and therefore average costs for each need an alternate definition. The ray average cost definition offers a way of dealing with and evaluating scale economies in this multiproduct environment.

Measurement and conceptual issues are also important in hospital cost studies. The nature of how the decision making unit is defined appears initially trivial. However, considering that sources of scale economies are typically thought of as shared resources, it is not clear whether the site or organisation should be the unit of analysis. In other contexts, the site level may be preferable, but in the UK, ‘hospital chains’ are not present as in other countries, in the sense of large organisations providing multiple hospitals across several cities. Therefore, in the subsequent analysis, the organisation level is used as the unit of analysis.

We also need to consider whether to define outputs as health or healthcare. Whilst there may be theoretical arguments for either, the practicalities dictate that healthcare is defined using outputs rather than health gains. There is simply not enough data on health improvements for each treatment. Having decided to use healthcare as the output, subsequent essays categorise healthcare by the number of cases, not patient-days. The additional costs associated with longer stays should have already been accounted for by HRG weighting. Counting bed days as outputs also has a distortionary effect where stay lengths are protracted. Longer stays may result from hospital inefficiency or capacity constraints in follow-on care rather than reflecting additional treatment. Consequently, subsequent chapters take weighted case numbers as the measure of output.

A degree of output aggregation is inevitable, given the vast heterogeneity of hospital treatments. In the UK, the best means of aggregation and dealing with casemix is to use HRG-based weightings. HRG codes include information on the treatment offered by the hospital and relevant elements of the patient's underlying health. HRGs also have a 'built-in' weighting schema as they are used for tariff payments. In contrast, alternate means of weighting, such as ICDs, may not adequately distinguish between outputs. Large numbers of patients with similar ICD codings may receive different treatments. Patient treatment may differ dependent on the subtleties of their need not captured by ICD codes or by the exercise of patient choice.

Additionally, there is a question of how to measure 'size' or scale. Most studies have followed Friedman (1955) and eschewed measuring size in terms of outputs due to the bias engendered by higher-output hospitals having both higher output and lower average costs. This thesis follows the approach of most prior studies by using bed numbers as the best substitute. Book value of financial statement assets offers an alternative, but valuation uncertainty and accounting practice mean this is as much of an abstraction as bed numbers.

Methodologically, parametric methods are preferred in this thesis. The attraction of parametric over non-parametric methods lies in the ability to account for measurement error and the relative robustness to outliers. Parametric methods also allow the imposition of desirable economic foundations or at least render these testable. Problems with the direct estimation of long-run cost functions can be addressed by taking an envelope of short-run costs. Both direct long-run estimation and envelope methods are considered in the remainder of this thesis. Regarding the choice of functional form, both translog or quadratic functional forms are used, as is appropriate for investigating scale economies which vary with size.

Chapter 2

Economies of Scale in English Hospitals: A Translog Estimation

2.1 Introduction

Hospital care consumes half of UK government health expenditure (Pett, Tristan et al., 2020). Much policy is devoted to identifying inefficiencies, but comparatively little on the degree of scale economies in the provision of hospital care. Quantifying the relation between size and cost is important to understand the financial effects of current policy. Recent policy changes in England aim to supply services at a regional, systemic level via 'Integrated Care Systems' (Ham, 2018). Though this policy framework is primarily concerned with coordinating different health and social care organisations, it has also helped to facilitate the centralisation of services. Smaller services hitherto distributed over several providers are increasingly consolidated within a larger hospital in the local area. The resulting fewer organisations can supply more healthcare services at scale than the preceding decentralised 'district general hospital' model.

Previous studies have suggested that, in certain circumstances, concentrating care in fewer larger institutions improves the *quality* of care (Begg et al., 1998). However, evidence concerning the effect on the *cost* of care is limited, despite a frequent assumption that larger-scale services ought to be able to realise scale economies (Goudie & Goddard, 2011). Where prior studies of scale and cost have been completed, they are now dated, and no studies considering all hospital activity have been successfully carried out using data for any of the UK nations. The English National Health Service ('NHS') provides an interesting case study, having a higher degree of concentration than most other healthcare systems. English hospitals are also typically larger than those in other countries, as measured by the number of beds, so they offer an opportunity to test a higher range than in other countries.

I address this gap in the literature by examining scale economies in English hospitals, using data from the English NHS's National Cost Collection exercise showing healthcare outputs and costs. This dataset is linked to summary wage data extracted from the national human resources database and capital costs from financial statements. Using this linked data, a translog long-run cost function is estimated. The translog form allows average costs to vary with output levels and therefore is suited for estimating over what range of outputs scale economies are positive.

The results show that there are unexploited scale economies. Scale economies are positive up to 1,200 beds, compared with the median hospital that had an average of 741 beds between 2013/14 and 2018/19. To offer a more rounded picture of the relationship between scale and average costs, two alternate measures of the limit of scale economies are also computed using output measures. Scale economies are exhausted at the level of 90,000 elective inpatients or 1,000,000 outpatients. For comparison, the median hospital completes 47,000 elective care episodes and 531,000 outpatient appointments. An additional expansion of activity at hospitals below this limit could be accomplished at a proportionately lower cost than those with activity at the higher end of the distribution. Scale economies were also found to be larger in lower wage areas outside of London and the South East of the country, and higher for non-elective admitted care relative to ambulatory or elective admissions.

This study updates our understanding of the limits of hospital scale economies, an area where the literature would benefit from updated estimates. It uses evidence from a country rarely studied but with a high amount of integrated data available. The additional data also allows the consideration of variation in input prices in the cost specification, which not all prior studies can do, relying on the assumption that centralised procurement or wage bargaining ensures hospitals face the same input prices.

2.2 Previous Literature

Previous literature uses various methods and data to estimate hospital economies of scale. There is an extensive literature using data envelopment analysis or stochastic frontier to measure departures from optimal production levels, including the efficient scale of operations (Giancotti et al., 2017). However, these studies typically report variation from the optimal scale as an index for the sector as a whole rather than estimating the minimum efficient scale. This approach makes it hard to find the point at which scale economies are expected to become negative, assuming they decline with size. This approach is followed in some parametric studies, for example Carey et al. (2015), who are more concerned with differences between specialist hospitals and general hospitals, but report positive scale economies up until the third quartile of their US dataset for general hospitals, and up to the median for the case of for profit generalist hospitals. Li & Rosenman (2001) use a parametric leontief cost function and find positive economies of scale over the sample, but do not report the relationship with size or output.

Studies that do report particular values for limits to scale economies tend to find limits that are low relative to the size of UK hospitals. Estimates have varied between 125 beds using a DEA approach on Greek data (Fragkiadakis et al., 2016), 230 beds (Azevedo & Mateus, 2014) in Portugal using parametric techniques, and 370 beds (Frech & Mobley, 1995) in the US, again using parametric methods. (Wilson & Carey, 2004) used non-parametric techniques and found no evidence of limits to scale economies using US data that covers hospitals up to a size of 400 beds. In contrast, the median English hospital had 740 beds during the fiscal year 2018-19, far above the limits found in previous studies. It is instructive to use data from this more consolidated setting to test whether results differ. To my knowledge, no estimation of hospital-level scale economies in the English NHS has been completed, though individual departmental level studies have been carried out (NHS Centre for Reviews and Dissemination, 1996). Freeman et al. (2020) find evidence of scale economies within inpatient care but do not consider care provided in other settings, such as outpatient care. They also do not report an estimated limit to scale economies. A previous attempt to measure scale economies in Scotland (Scott & Parkin, 1995) was unable to estimate a cost function due to weaknesses in the data or accounting methods used at the time. Since Scott and Parkin's study, developments in the availability of data now make estimation feasible.

Older papers, now likely affected by technological change since their publication, offer varying assessments of scale economies. (Dranove, 1998) used discharges as a measure of scale, finding a limit at 10,000 discharges. Other papers typically reported contemporaneous scale economy estimates present rather than noting limits. Results varied from positive returns to scale using countrywide US data (Berry, 1967), constant returns to scale using data from North Carolina (Conrad & Strauss, 1983), to slight diseconomies using Californian data (Vita, 1990). No evidence for long-run scale economies was found using Pennsylvanian data (Lave & Lave, 1970) or data from Queensland (Butler, 1995, ch. 7). Consequently, a study providing updated estimates from a country not often studied is a useful addition to the literature.

There is also a rich associated literature exploring factors affecting hospital costs and productive efficiency. Asmild et al. (2013) use a DEA production function analysis to find that productive efficiency and optimal scale can vary according to the level of urbanisation. Additional requirements of teaching can also be associated with higher costs amongst hospitals involved in medical training (Culyer et al., 1978; Sloan, Frank A et al., 1983), as can the requirement to produce a broader range of specialised services (Farsi & Filippini, 2008), and local patient health (Clark, 2012). Other factors influencing cost include the degree of competition where this is applicable (Fournier & Mitchell, 1992), quality of care outcomes (Gutacker et al., 2013), the degree of surplus bed capacity (Keeler & Ying, 1996), waiting list times (Siciliani et al., 2009), and the administrative model of the hospital (Weaver & Deolalikar, 2004).

2.3 Data

2.3.1 Cost Data

Acute care cost and output data from the NHS National Cost Collection exercise was accessed through the NHS England Reference Costs website (NHS Improvement, 2019). Under this exercise, each NHS healthcare provider is mandated by NHS England to provide information on costs and healthcare outputs. Similar exercises are carried out in the devolved administrations of Scotland, Wales and Northern Ireland. However, there are methodological differences in the data collection exercise between nations and differences in the organisational structure for admitted healthcare. Consequently, to allow observations to be compared, data was only taken from the largest nation (England).

To generate this data, providers take total expenditure from the same systems that generate published financial statements and remove income and expenditure items unrelated to patient care. Exclusions include training, research, private patient treatment, miscellaneous activities such as joint ventures, and accounting adjustments. The remaining expenditure is then allocated to individual treatment episodes according to known information about that episode. For example, patients spending a certain number of days in an inpatient ward are allocated a proportionate amount of annual inpatient nursing expenditure. The data details expenditure split according to several dimensions, including the medical or surgical specialty, the output type, and the diagnosis or treatment requirement described by the Healthcare Resource Group (‘HRG’) code, see Chapter 1. Healthcare outputs are categorised in the dataset according to where the patient was treated or how the patient was admitted. These categories form the basis of the split of output types in the multi-product cost function estimated, which includes output categories such as admitted patient care (elective and non-elective episodes), outpatient appointments, emergency department attendances, critical care episodes and diagnostic services.

The dataset has an inherent advantage over expenditure data based on financial statements, as research, teaching, private patient and accounting adjustments are excluded automatically by the collection method. Expenditure recorded in this dataset is consequently closer to the underlying cost of healthcare than summary expenditure recorded in organisation financial statements.

Outputs are restricted to acute care only by excluding irrelevant services and organisations. As noted in Chapter 1, a hospital is defined as an organisation that provides elective or non-elected acute inpatient care. In the UK, the legal entities that provide such care may operate more than one hospital, though none would have so many as to be reasonably thought of as a 'hospital chain'. For conciseness, such organisations are referred to as 'hospitals'. Non-acute or community services carried out by hospital providers are omitted, as are the costs of specialist medicines prescribed in the hospital but taken at home. Specialised activities such as rehabilitation, renal dialysis, and cystic fibrosis treatment are also excluded due to their more complex costing methods and associated cost structures. This dataset was then filtered to include only organisations providing acute secondary care, excluding mental health care providers, community care providers and specialist organisations, as discussed in Chapter 1. These excluded organisations are likely to have different technologies and cost profiles than a typical acute organisation and are not appropriate to include in the analysis. Specialist organisations excluded from the analysis are listed in Appendix A.

One organisation was also excluded from the dataset because it did not use the national NHS staff records system and consequently lacked wage data. Another organisation had one observation removed where diagnostic services output was zero, likely a recording error. These exclusions left a panel of comparable acute care organisations and outputs recognisable as hospital acute care. The final dataset is a panel of 141 acute organisations with at least one year of output data between the 1st of April 2013 and the 31st of March 2019, the most recent data available as of October 2022. In the period covered by the panel, the number of organisations declined from 138 in 2013/14 to 127 in 2018/19. The reduction is the result of organisations merging or closed either due to poor clinical standards, unsustainable finances, policy changes about local healthcare supply, or some combination of these factors.

Each output category is defined by the type of healthcare provided and the setting in which it takes place. Elective and Non-Elective Episodes count periods of care under a named senior doctor where the patient is admitted to the hospital for at least one overnight stay, differing according to whether the admission was planned (Elective) or unplanned (Non-Elective). Outpatient Attendances count attendances at a hospital clinic where the patient is not required to use a bed, typically only staying in the clinic for a short consultation. Emergency Department attendances count the number of attendances in the Emergency Department. Where the patient requires a subsequent admission to the main hospital, an additional non-elective episode would be recorded for the same patient. Critical Care episodes record the times a patient requires an intensive care stay. Finally, Diagnostic Services comprise a variety of tests and imaging, from blood tests to x-rays and CT scans.

Table 2.1 sets out summary statistics for output levels by the type of healthcare provided. There is a right skew to all output types, with a long right tail of higher output organisations. There are a greater number of lower-output organisations, demonstrated by the median being lower than the mean across all output types. The number of admitted care outputs, defined as Elective and Non-Elective care episodes and Critical Care episodes, is much lower than the number of Outpatient healthcare outputs.

Table 2.1: Annual Hospital Output Descriptive Statistics (April 2013 - March 2019)

Healthcare Output Type	**Mean**	**Median**	**SD**	**Q1**	**Q3**
Critical Care Episode	19,091	13,592	16,362	8,484	22,592
Diagnostic Services	2,973,184	2,693,514	2,117,147	1,541,434	3,876,838
Elective Admission	50,422	45,667	25,159	31,504	64,240
Emergency Dept Attendance	129,432	112,520	64,773	86,049	153,076
Non-Elective Admission	69,921	62,642	32,438	47,117	87,996
Outpatient Attendance	579,893	520,824	285,127	369,222	726,010

N = 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19] hospitals

Table 2.2 sets out the shares of total expenditure incurred under each output type in 2018/19, the most recent year in the dataset. Though individual outpatient average costs are low, the large number of appointments carried out means they are 27% of total costs. Analogously, the low cost of most diagnostic services means that though raw output numbers are high, total costs for these services are comparatively low. The majority of expenditure is incurred in three main categories of output: Non-Elective Episodes, Outpatient Attendances and Elective Episodes, comprising 79% of total costs. Emergency Department Attendances, Critical Care, and Diagnostic Services make up a lower proportion of the total cost.

Table 2.2: Total Output Costs (£m, 2018/19 only)

Activity Type	**Cost**	**Proportion**
Non-Elective Admission	12,706	33%
Outpatient Attendance	10,385	27%
Elective Admission	7,427	19%
Emergency Dept Attendance	3,220	8%
Critical Care Episode	2,965	8%
Diagnostic Services	1,972	5%
Total	**38,675**	**100%**

N = 550,032,118 healthcare outputs

Summary descriptive average cost statistics for these output categories are set out in Table 2.3, using observations from all years in the dataset. The most costly output types are admitted inpatient episodes, as these outputs involve ward stays, possibly including surgery. Non-Elective Episodes have more variation in cost as these cases can involve greater treatment requirements at the upper end of the distribution. Average costs for Outpatient and Emergency Department Attendances are much lower as these types of output are much shorter in duration and do not involve a hospital stay. If an emergency department attendance requires subsequent admission to the hospital, then healthcare provided after this point is categorised under 'Non-Elective Episode'. The standard deviations show significant variation in average costs and a long right tail of higher episodic costs. This large variation arises due to differences in acuity and required treatment. Admitted care can range from a day's stay for a minor fracture with no required interventions to very complex cranial or vascular repair surgery. This cost variation requires us to weight outputs to account for cost differences even within these output groups.

Table 2.3: Average Costs of each Activity Type (£s, April 2013 - March 2019)

Activity Type	Mean	SD	Median	Q1	Q3	95%
Critical Care Episode	1,067	574	995	600	1,431	2,021
Diagnostic Services	5	22	1	1	2	14
Elective Admission	1,259	2,028	642	364	1,288	4,697
Emergency Dept Attendance	156	87	143	99	193	313
Non-Elective Admission	1,552	2,119	738	401	2,021	4,938
Outpatient Attendance	122	98	106	69	151	269

N = 3,049,910,728 healthcare outputs

2.3.2 Weighting

Within the dataset's output categories, there is considerable heterogeneity of care and expected resource requirements. For example, elective care treatment can vary from minor fractures to more complex reconstruction surgery. In the analysis below, outputs are re-weighted within each category using an HRG-based index that adjusts output levels to account for this heterogeneity, basing the weights on HRG average costs in 2019/20. As the number and specificity of HRGs tends to increase over time, the latest financial year in the dataset (2019/20) was used to ensure the maximum number of HRG matches in prior years. To remove a potential source of endogeneity, the year 2019/20 was subsequently excluded from the dataset. More than 95% of outputs can be described using 2019/20 HRG codes. This approach was not possible where codes changed between the recorded output and 2018/19. In these cases, the successor code was used where there was a 1:1 relationship between the old and new code. Assigning these codes took the percentage of outputs expressible in 2019/20 codes to more than 99% of the total output. Where there was no 1:1 relationship to a 2019/20 code, the weighting from the latest year where that code existed was used. Consequently, each output was assigned a weight that made it comparable to other outputs.

The quantity of each healthcare output was multiplied by a weighting ratio to construct the weighted output measure. This ratio was constructed as the 2019-20 average cost for that HRG and output category divided by the average cost for that output category across all HRGs. For example, in 2019/20, the average cost of an elective care episode coded "FF31D - Complex Large Intestine Procedures 19 years and over with CC Score 0-2" was £8,205. This figure was divided by the average cost of an elective care episode across all HRGS in 2019/20, which was £3,877. Consequently, that procedure receives a weight of 8205/3877 = 2.1 and outputs with this coding are multiplied by 2.1 to reflect the relative complexity of this particular procedure. This approach results in healthcare outputs expressed in terms of a numeraire good equal to the average cost for that combination of HRG and output category in 2019-20. The weighting index isolates the variation in cost explicable by differences in the type of healthcare supplied. This approach thereby makes comparable organisations which treat many patients but for comparatively minor ailments and those treating fewer patients but those requiring more complex care.

This weighting of outputs was necessary to deal with the heterogeneity in healthcare goods and patient comorbidities. Early literature surveyed by Mann & Yett (1968) and Hefty (1969) accounted for heterogeneity by splitting outputs into categories, such as admitted and outpatient care, or by modelling outputs as bed days rather than discrete episodes of healthcare provided. These categorisations dealt with some variation across output categories but failed to recognise the considerable differences in output within such categories. Even where care outputs fall within the same broad category such as 'admitted care', there can be large differences between costs associated with the different production technologies used. For example, the resource requirements to provide a week long stay in a medical ward may be less than for a week long stay in a surgical ward, where the additional requirements of surgery would have to be considered. Recognising this limitation, more recent literature (Preyra & Pink, 2006) accounts for differences in healthcare outputs and technologies by using an index or weight based on the expected resource requirement. Using HRG weights, healthcare output heterogeneity within each category is accounted for, avoiding the issues in earlier literature.

2.3.3 Input Prices

2.3.3.1 Workforce and Wage Data

Aggregated staff numbers and average wages are taken from the NHS Electronic Staff Record ('ESR'), accessed via NHS England's workforce statistics group. The ESR database contains, for all NHS staff, data on employing organisation, profession, level of seniority, headcounts, working hours, and wages. Though wage rates are set nationally, there is some variation in individual wage rates according to seniority, being in high-cost areas proximate to London, and additional responsibilities, including payments for working unsocial hours. An average wage rate for each organisation's doctors, nurses and administrative staff was computed by dividing total wages by the sum of full-time equivalent staff.

Summary information on average earnings rates is presented in Table 2.4, setting out the average organisational wage payments split by type of staff. Figures are adjusted to account for part-time staff and are presented as 'full-time-equivalent' payments. Payments are presented in cash terms, without adjustment for inflation. Over the period considered, non-medical wages were stable in cash terms, with little growth. The lack of wage growth in the period resulted from a public sector pay cap imposed by the government. This cap was part of efforts to reduce public spending following the 2007/08 financial crisis. Under the cap, annual pay awards were restricted to 1% until 2017/18. The pay cap affected all staff other than doctors, whose pay was not subject to the cap and is negotiated in a separate process.

Table 2.4: Earnings by Staff Group (£s)

Staff	**Measure**	**2012/13**	**2013/14**	**2014/15**	**2015/16**	**2016/17**	**2017/18**	**2018/19**
Senior Doctor Wages	Mean	101,104	102,802	104,119	104,527	105,719	106,643	107,538
	SD	6,411	6,445	6,513	6,505	6,956	7,340	7,450
Training Grade Doctor Wages	Mean	46,950	46,072	46,582	46,570	47,567	48,799	50,210
	SD	2,202	2,927	2,442	2,441	2,290	2,500	2,734
Nurse Wages	Mean	34,820	34,966	35,086	35,191	35,393	35,782	36,666
	SD	1,893	1,871	1,862	1,795	1,780	1,848	1,926
Healthcare Assistant Staff Wages	Mean	20,684	20,840	20,920	21,109	21,227	21,414	22,298
	SD	1,612	1,602	1,595	1,511	1,506	1,540	1,596
Administrative Staff Wages	Mean	29,396	29,597	29,923	29,992	30,258	30,817	31,808
	SD	4,723	4,699	4,772	4,519	4,448	4,638	4,622

Senior Doctors include consultant and associate specialists, Training Grade Doctors include registrars and lower grades

Geographical variation in wage rates can be seen in Figures 2.1 - 2.4, which plot wage quintiles for senior doctors, junior doctors, nurses and administrative staff across England. Nurses and administrative staff are given higher wages in London to reflect the higher local price level, as shown in Figures 2.3 - 2.4. Medical pay uplifts for London are considerably smaller, so variation in medical wage rates is more evenly distributed across the country.

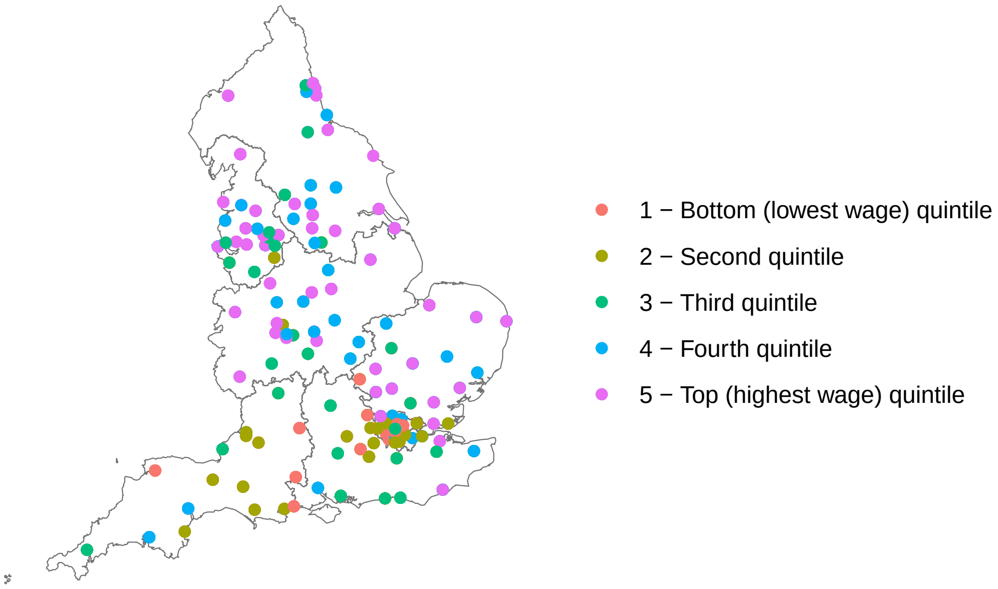
● 1 − Bottom (lowest wage) quintile
● 2 − Second quintile
● 3 − Third quintile
● 4 − Fourth quintile
● 5 − Top (highest wage) quintile

Figure 2.1: Geographical Variation in Senior Doctor Wages

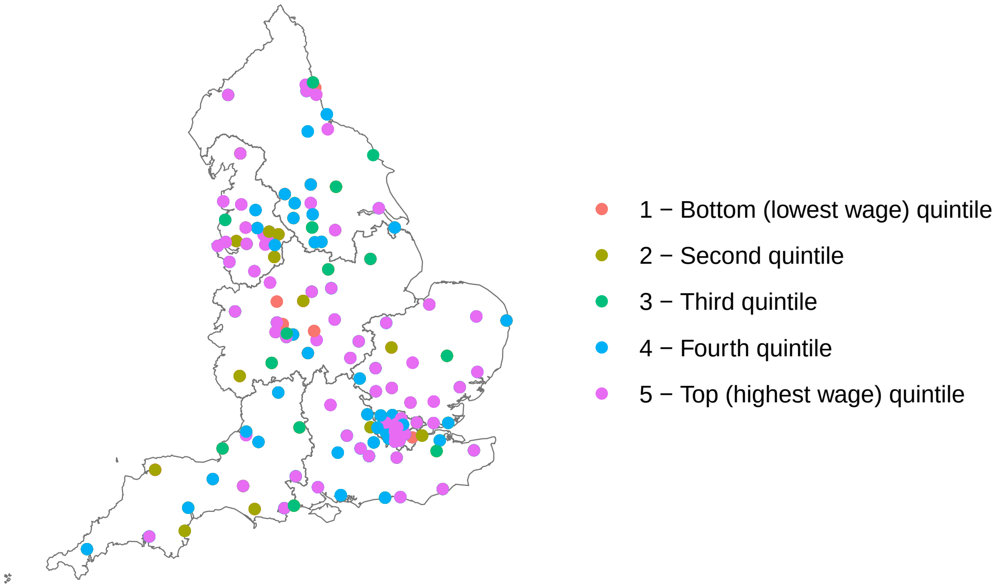
● 1 − Bottom (lowest wage) quintile
● 2 − Second quintile
● 3 − Third quintile
● 4 − Fourth quintile
● 5 − Top (highest wage) quintile

Figure 2.2: Geographical Variation in Training Grade Doctor Wages

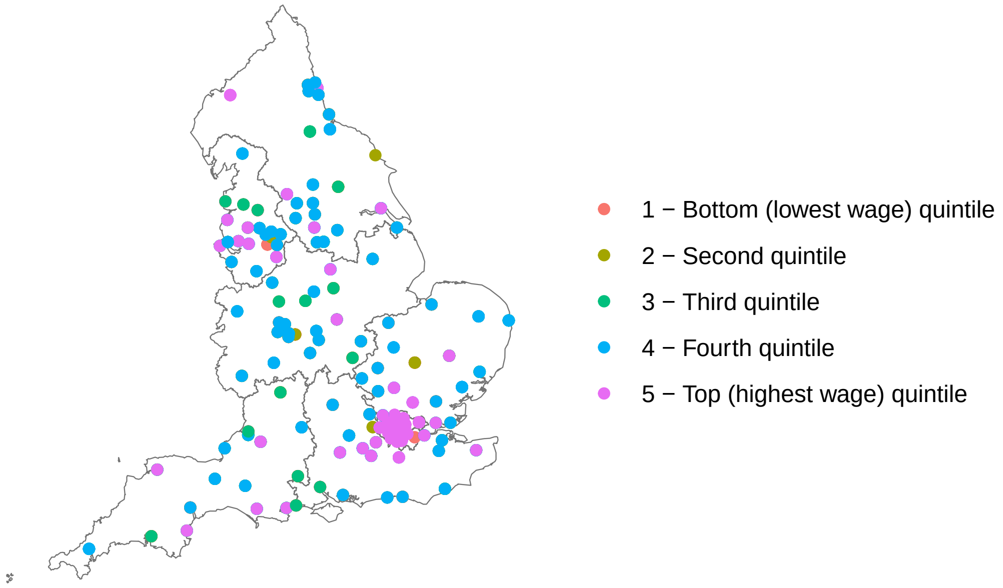
1 − Bottom (lowest wage) quintile
2 − Second quintile
3 − Third quintile
4 − Fourth quintile
5 − Top (highest wage) quintile

Figure 2.3: Geographical Variation in Nurse Wages

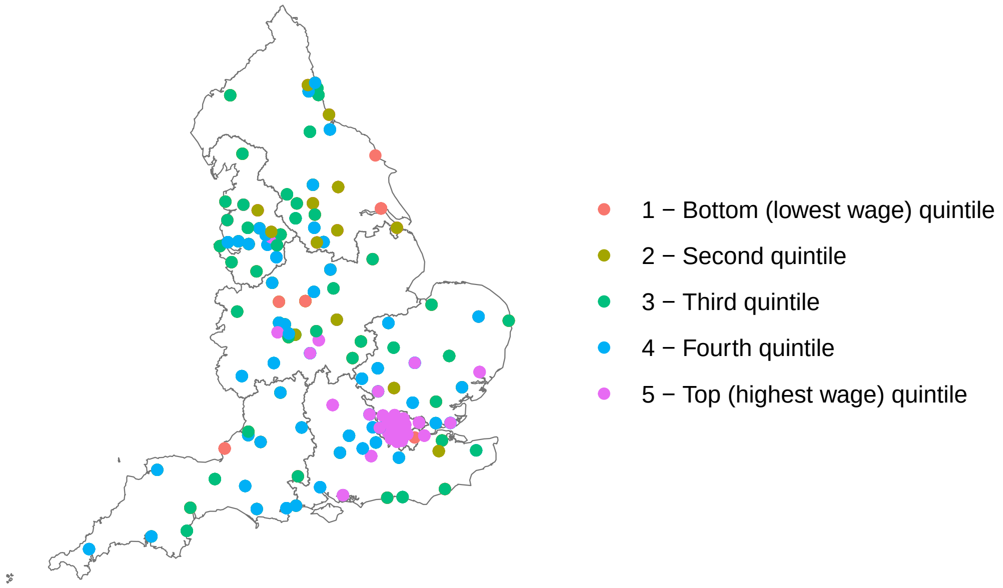
1 − Bottom (lowest wage) quintile
2 − Second quintile
3 − Third quintile
4 − Fourth quintile
5 − Top (highest wage) quintile

Figure 2.4: Geographical Variation in Administrative Staff Wages

2.3.3.2 Capital Costs

An average weighted cost of capital is calculated by obtaining financing costs and asset valuations from each organisation's financial accounting statements. The cost of capital is defined as the ratio of total financing costs to the value of revalued plant, property and equipment. Summary statistics for this measure are set out in Table 2.5. Though most financing of English hospitals is provided through the government, there are variations in the effective rate paid, either due to variations in the public dividend rate charged to organisations or the use of non-state financing by the respective organisation. Financing costs have remained broadly constant over the period considered, with little change to the mean or median financing cost or the variation observed between organisations.

Table 2.5: Weighted Average Cost of Capital (%)

Statistic	**2013/14**	**2014/15**	**2015/16**	**2016/17**	**2017/18**	**2018/19**
Mean	4.67%	4.73%	4.77%	4.92%	3.74%	3.84%
Median	4.20%	4.15%	4.19%	4.27%	3.03%	3.06%
SD	1.84%	1.95%	1.91%	2.09%	2.05%	2.19%
Q1	3.39%	3.36%	3.43%	3.41%	2.56%	2.53%
Q3	5.37%	5.55%	5.87%	5.84%	4.69%	4.31%
N (hospitals)	**140**	**136**	**135**	**135**	**133**	**130**

Capital costs are predominantly payments to central government in exchange for historically gifted assets ('Public Dividend Capital') but may also be augmented by commercial borrowing, including lease payments. A weighted average cost of capital is derived by dividing total interest by total relevant assets

2.3.3.3 Collinearity, Wage Rates, and Input Price Selection

In the English NHS, wage rates are set by central government rather than individual hospitals. Consequently, some staff categories exhibit multicollinearity where they are part of the same wage-setting framework. For example, administrative, nursing and healthcare assistant wages are determined by the same framework, as are senior and training grade doctors. Whilst multicollinearity will not bias the estimated coefficients, it does mean that they are less accurate, with inflated standard errors, and could be more sensitive to small changes in the underlying data.

To quantify the effect of multicollinearity amongst wage variables, I conduct a principal component analysis of all input variables, noting the number of clusters and the correlation of each input variable with the principal components generated. Five components explained more than 95% of the variance. Figure 2.5 illustrates how each principal component correlates with each input price. Component 1 comprises all staff types covered by the same pay agreement (administrative, nursing and healthcare assistant staff), while component 2 is predominantly made up of medical staff, also governed by a common pay framework. The capital rate measure is the dominant item within component 3. In the subsequent analysis, input prices shown to be highly collinear were dropped, as they do not add much additional information to the model. For this reason, and to retain a parsimonious model, subsequent analyses include the nursing staff wage variable, the training grade doctor variable and the capital rate measure as input prices in the main model. Nursing staff are also the most numerous staff members in the first pay framework category, as are training grade doctors in the second.

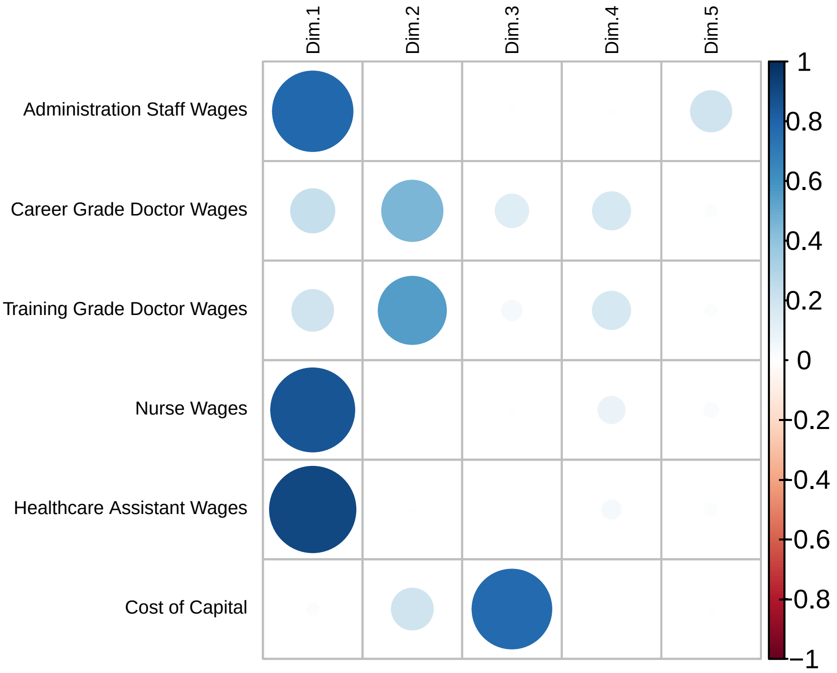

Figure 2.5: Correlation between Input Prices and their Principal Components

2.3.4 Data Structure

The dataset for estimation was constructed from the cost and input price data above, with the addition of a dummy variable for teaching hospital status. There are several possible definitions for what constitutes a teaching hospital, with no universally recognised demarcation in the UK. In the analysis below, organisations are classified as teaching hospitals where they serve a medical school, have membership of the Association of UK University Hospitals, or indicate teaching hospital status on their website. This definition covered 51 organisations between 2013/14 and 2016/17 and 52 in 2017/18 and 2018/19.

The final dataset was an unbalanced panel covering 6 years, with 141 hospitals being in operation for at least some of that period. The number of hospitals declined from 138 to 138 between 2013/14 and 2018/19 due to reorganisations, the creation of new hospital organisations from mergers, and closures. There were 798 total observations (hospital-year pairings) in the final panel.

2.4 Method

2.4.1 Defining Scale Economies

As discussed in Chapter 1, it is more complex to define economies of scale in the multiproduct case than in the case of single good production. The definition of scale economies is taken from that chapter, using ray average costs when evaluating average and marginal cost ratios. The scale economy measure becomes:

$$\begin{aligned} S &= \frac{AC}{MC} \\ &= \frac{\left(\frac{C(w,y)}{y}\right)}{\left(\frac{dC(w,y)}{dy}\right)} \\ &= \frac{dy}{dC(w,y)} \cdot \frac{C(w,y)}{y} \\ &= \frac{d\ln y}{d\ln C(w,y)} \\ &= \frac{1}{\epsilon_{C,y}} \end{aligned} \tag{2.1}$$

To resolve the issue of how changes to the multiproduct vector **y** are treated, I make the standard assumption that increases in inputs occur proportionately from current output levels (M. A. Fuss & Waverman, 1981). In this case, total scale economies can be expressed as the inverse of the sum of individual total cost elasticities for each y_i, $\epsilon_{C,y_i} = \partial lnC(w,y)/\partial lny_i$:

$$\begin{aligned} S &= \frac{1}{\sum_i^m \epsilon_{C,y_i}} \\ &= \frac{1}{\sum_i^m \partial lnC(w,y)/\partial lny_i} \end{aligned} \tag{2.2}$$

Overall scale economies can then be evaluated as the inverse of the sum of the individual total cost elasticities, where each elasticity is calculated using total cost.

2.4.2 The Translog Cost Function

Studies that attempt to determine specific ranges for scale economies typically use parametric methods to estimate a cost function, usually in a 'flexible form' such as a transcendental logarithmic or quadratic specification. The cost function estimated in this chapter is a panel version of the multiproduct translog functional form outlaid in Chapter 1, as follows:

$$\begin{aligned}\ln\, C_{ht}(y,w) =& \alpha_0 + \sum_i^m \alpha_i \ln\, y_{iht} + \sum_k^n \beta_k \ln\, w_{kht} + \frac{1}{2}\sum_i^m \sum_j^m \delta_{ij} \ln\, y_{iht} \ln\, y_{jht} \\ &+ \frac{1}{2}\sum_k^n \sum_l^n \gamma_{kl}\, ln\, w_{kht}\, ln\, w_{lht} + \sum_i^m \sum_k^n \rho_{ik}\, \ln\, y_{iht}\, ln\, w_{kht} + \epsilon_{ht}\end{aligned} \tag{2.3}$$

y_i is the amount of the ith output produced (of m total outputs), w_i is the ith input price (of n inputs), and α_i, β_k, δ_{ij}, $\gamma_{k,l}$, and $\rho_{i,k}$ represent coefficients estimated in the model. Hospital index h and time period index t reflect the panel data nature of the dataset. The error term $\epsilon_{ht} = \tau_h + v_{ht}$ comprises a term τ_h representing hospital specific effects and a random noise term v_{it}. For reasons of clarity, time and individual subscripts are omitted from this point on in the chapter.

The random effects specification assists in controlling for unobserved factors associated with specific hospitals. The alternate fixed effects specification would perform poorly given the relatively short time period observed and the fact that output and price variables are relatively time-invariant. Consequently, including fixed effects would give rise to multicollinearity problems, and a random effects model is preferred.

To measure the effects of changes in output quantities on overall costs, point cost elasticities for a specific output i are computed according to the partial derivative of the total (log) cost equation with respect to log y_i:

$$\epsilon_{C,y_i} = \frac{\partial \ln C(w,y)}{\partial \ln y_i} = \alpha_i + \sum_j^m \delta_{ij} \ln y_j + \sum_k^n \rho_{ij} \ln w_k \tag{2.4}$$

Following on from the definition of total scale economies given in Equation (2.2), the inverse of each individual total cost-output elasticity scale economy measurement for y_i is:

$$S_i = \frac{1}{\epsilon_{C,y_i}} = \frac{1}{\alpha_i + \sum_j^m \delta_{ij} \ln\, y_j + \sum_k^n \rho_{ij} \ln\, w_k} \tag{2.5}$$

Total scale economies $\mathbf{S}$ are the sum of these individual cost elasticities:

$$\begin{aligned}\mathbf{S} &= 1/\sum_{i}^{m} \epsilon_{C,y_i} \\ &= \frac{1}{\sum_i^m \left(\alpha_i + \sum_j^m \delta_{ij} \ln y_j + \sum_k^n \rho_{ij} \ln w_k\right)}\end{aligned} \tag{2.6}$$

Where $\mathbf{S} > 1$, economies of scale prevail, where $\mathbf{S} = 1$ average costs are constant with changes in output, and where $\mathbf{S} < 1$, diseconomies are present.

2.4.3 Other Variables and Estimation

In addition to the output measures, input prices and cost products, a dummy variable for teaching costs is also included. As discussed in Chapter 1, hospitals incur additional costs because they are required to teach medical students. This expenditure is independent of possible confounders like case complexity in various settings (Culyer et al., 1978; Sloan, Frank A et al., 1983). According to the dataset methodology, hospitals should exclude teaching costs during collection. However, as it is difficult for hospitals to identify the costs of staff time spent teaching, a teaching hospital dummy is included as an explanatory variable in the estimation, to absorb variation in teaching-related expenditure.

Costs may be affected by the quality of care offered. However, quality is difficult to measure and account for in one variable. Whilst many different quality measures are collected for each hospital, most are partial and also affected by unobserved factors like the underlying health of the local population. The broadest measure that attempts to take account of local population health would be the Summary Hospital-level Mortality Indicator (SHMI). The SHMI is a ratio of observed and expected hospital deaths occurring either in the hospital or shortly after discharge. The SHMI is not a direct measure of quality of care. However, it could be used as a proxy measure to isolate the variation in cost attributable to differences in care quality.

The translog cost function is estimated using a random effects estimator with heteroskedasticity-robust standard errors, using the plm package (Croissant & Millo, 2008) in R (R Core Team, 2022). After log transformation, output and price variables are also centred by subtracting the sample means. Output and price variables are centred to deal with multicollinearity amongst these variables. Centring these variables does not improve the overall variance and fit of the model but reduces the size of correlations between the output and price variables and their cross-products (Iacobucci et al., 2016).

After the translog cost function is estimated, it is checked for consistency with the desired theoretical properties. To confirm that the estimated cost function is plausible, the goodness of fit is investigated, along with visualisations of average and marginal cost curves. Finally, scale economies are computed according to Equation (2.5), and the relation between size and scale is explored.

2.5 Results

Descriptive statistics for the output categories are set out in Table 2.6 below. In the estimation, values are centred at the mean, but are presented prior to the transformation in the table.

Table 2.6: Descriptive Statistics

Non-Centred Values	**N = 798**
A&E Attendance	
Mean (SD)	127,483 (60,463)
Median (IQR)	112,013 (85,640, 152,096)
Range	33,545, 447,391
Critical Care Episode	
Mean (SD)	19,988 (19,585)
Median (IQR)	12,709 (8,016, 21,496)
Range	2,001, 120,697
Diagnostic Services	
Mean (SD)	4,297,133 (2,508,860)
Median (IQR)	3,777,083 (2,642,614, 5,255,448)
Range	333,272, 26,055,137
Elective Inpatient Episodes	
Mean (SD)	51,604 (25,862)
Median (IQR)	46,910 (31,734, 64,884)
Range	9,333, 165,462
Non-Elective Inpatient Episodes	
Mean (SD)	71,745 (34,026)
Median (IQR)	65,122 (47,133, 90,191)
Range	16,051, 231,458
Outpatient Attendances	
Mean (SD)	591,701 (293,106)
Median (IQR)	530,505 (377,434, 738,400)
Range	134,484, 2,053,768
Cost of Capital	
Mean (SD)	4.47 (2.06)
Median (IQR)	3.94 (3.04, 5.36)
Range	0.67, 13.91
Nursing Wages	
Mean (SD)	35,501 (1,930)
Median (IQR)	34,923 (34,260, 35,970)

Table 2.6: Descriptive Statistics

Non-Centred Values	**N = 798**
Range	32,703, 42,253
Training Grade Doctor Wages	
Mean (SD)	47,630 (2,828)
Median (IQR)	47,431 (45,632, 49,543)
Range	40,077, 55,880
Teaching Hospital Organisation	
0	523 (66%)
1	275 (34%)

Estimation results for the cost function are set out in Table 2.7.

Table 2.7: Estimation results - Directly Estimated Translog Form

Term	**Estimate**	**Std. Error**
Intercept	19.356	(0.010)***
A&E Attendance	0.090	(0.015)***
Critical Care Episode	0.118	(0.011)***
Diagnostic Services	0.048	(0.008)***
Elective Inpatient Episode	0.303	(0.023)***
Non-Elective Inpatient Episode	0.251	(0.018)***
Outpatient Appointment	0.150	(0.021)***
Training Grade Doctor Wages	0.369	(0.068)***
Nurse Wages	0.504	(0.115)***
Cost of Capital	0.001	(0.008)
0.5 * A&E Attendance * A&E Attendance	-0.088	(0.049)
A&E Attendance * Critical Care Episode	-0.017	(0.032)
A&E Attendance * Diagnostic Services	0.059	(0.031)
A&E Attendance * Elective Inpatient Episode	-0.126	(0.073)
A&E Attendance * Non-Elective Inpatient Episode	0.128	(0.051)*
A&E Attendance * Outpatient Appointment	-0.057	(0.068)
0.5 * Critical Care Episode * Critical Care Episode	-0.030	(0.026)
Critical Care Episode * Diagnostic Services	-0.031	(0.018)
Critical Care Episode * Elective Inpatient Episode	0.037	(0.050)
Critical Care Episode * Non-Elective Inpatient Episode	0.022	(0.038)
Critical Care Episode * Outpatient Appointment	0.078	(0.046)
0.5 * Diagnostic Services * Diagnostic Services	-0.042	(0.019)*
Diagnostic Services * Elective Inpatient Episode	-0.020	(0.047)
Diagnostic Services * Non-Elective Inpatient Episode	0.041	(0.037)
Diagnostic Services * Outpatient Appointment	0.040	(0.044)
0.5 * Elective Inpatient Episode * Elective Inpatient Episode	0.188	(0.146)
Elective Inpatient Episode * Non-Elective Inpatient Episode	-0.024	(0.084)

R^2 = 0.9935 adj R^2 = 0.993 AIC = -2,631 BIC = -2,368 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

Table 2.7: Estimation results - Directly Estimated Translog Form

Term	**Estimate**	**Std. Error**
Elective Inpatient Episode * Outpatient Appointment	-0.075	(0.112)
0.5 * Non-Elective Inpatient Episode * Non-Elective Inpatient Episode	-0.145	(0.091)
Non-Elective Inpatient Episode * Outpatient Appointment	-0.009	(0.077)
0.5 * Outpatient Appointment * Outpatient Appointment	0.008	(0.127)
A&E Attendance * Training Grade Doctor Wages	0.596	(0.242)*
Critical Care Episode * Training Grade Doctor Wages	0.352	(0.156)*
Diagnostic Services * Training Grade Doctor Wages	-0.077	(0.152)
Elective Inpatient Episode * Training Grade Doctor Wages	-0.215	(0.345)
Non-Elective Inpatient Episode * Training Grade Doctor Wages	-0.279	(0.285)
Outpatient Appointment * Training Grade Doctor Wages	-0.299	(0.337)
A&E Attendance * Nurse Wages	0.032	(0.297)
Critical Care Episode * Nurse Wages	-0.229	(0.179)
Diagnostic Services * Nurse Wages	-0.256	(0.180)
Elective Inpatient Episode * Nurse Wages	0.881	(0.479)
Non-Elective Inpatient Episode * Nurse Wages	-0.488	(0.345)
Outpatient Appointment * Nurse Wages	0.454	(0.459)
A&E Attendance * Cost of Capital	-0.023	(0.028)
Critical Care Episode * Cost of Capital	0.053	(0.016)***
Diagnostic Services * Cost of Capital	0.028	(0.018)
Elective Inpatient Episode * Cost of Capital	-0.028	(0.035)
Non-Elective Inpatient Episode * Cost of Capital	-0.017	(0.029)
Outpatient Appointment * Cost of Capital	-0.035	(0.037)
0.5 * Training Grade Doctor Wages * Training Grade Doctor Wages	-2.520	(1.555)
Nurse Wages * Training Grade Doctor Wages	-3.028	(1.780)
0.5 * Nurse Wages * Nurse Wages	7.028	(3.238)*
Cost of Capital * Training Grade Doctor Wages	-0.024	(0.126)
Cost of Capital * Nurse Wages	0.239	(0.171)
0.5 * Cost of Capital * Cost of Capital	0.060	(0.019)**
Teaching Hospital Organisation	0.011	(0.016)

R^2 = 0.9935 adj R^2 = 0.993 AIC = -2,631 BIC = -2,368 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

2.5.1 Regression Results

All non-interaction terms were statistically significant, except for the cost of capital variable and the teaching hospital dummy. The exponent of the intercept

(£254,841,439) is comparable with the mean total hospital cost (£297,942,149), as expected, with output and price variables centred. Each non-interacted output coefficient was statistically significant and shows, in the absence of interaction effects, the proportionate increase in total cost from a 1% increase in that activity. The outputs with the most impact on total cost were elective and non-elective inpatients, with outpatient appointments and critical care episodes following and A&E attendances having the lowest impact on cost. Effects ranged from a 1% increase in elective inpatient activity increases total cost by 0.30%, absent interactions, to a 1% increase in A&E attendances increasing total costs by 0.09%. Of the input variables, nursing wages and medical wages were both statistically significant. Absent interaction effects, a 1% increase in medical wages increased total costs by 0.37%, the equivalent for nursing wages being 0.5. The constructed cost of capital variable was not statistically significant from zero.

Within the interaction terms, the interaction between A&E attendance and non-elective inpatients was positive and statistically significant. A 1% increase in A&E attendances (ceteris parabis) increases the cost elasticity with respect to non-elective inpatients by 0.13 percentage points. The self-cross product of diagnostic services is also statistically significant and negative, with additional percentage point increases in diagnostic service output associated with a -0.04% reduction in the cost elasticity for diagnostic services, suggesting declining marginal costs as output increases. Interactions between medical wages and A&E attendances were statistically significant, with a 1% increase in medical wages raising the cost elasticity of A&E attendances by 0.6%. A similar relationship was also the case between medical wages and critical care episodes, with a 1% increase in medical wages raising the cost elasticity of critical care episodes by 0.35%. A relationship between critical care episodes and the cost of capital was noted, with a 1% increase in the cost of capital increasing the cost elasticity of critical care episodes by 0.05%, perhaps due to the capital intensity of critical care. Finally, the self-cross products for the cost of capital variable and that of nursing wages were also statistically significant, which is expected, as increases in input prices would lead to greater effects of subsequent increases on total cost.

2.5.2 Consistency with Desired Theoretical Properties

The resulting cost function was evaluated according to the desirable cost function properties in Chapter 1.

1. **Non-negativity** — using log-transformed variables on the right hand side of the specification means that the underlying relationship between cost and the output and price variables will always be positive. This feature is an inherent

property of the translog specification that will be true irrespective of the model fit.

2. **Cost should be non-decreasing in input prices** — To check this condition, Shephard's Lemma $\frac{\partial C(w,y)}{\partial w_k}$ can be used to yield the derived demand function $x_k(x,y)$ for each input, differentiating the cost function with respect to the input price. The resulting formula gives us the cost-minimising input quantities for each input. For the multiproduct translog equation, this becomes:

$$\begin{aligned} x_k(w,y) &= \frac{\partial C(w,y)}{\partial w_k} \\ &= \frac{\partial \ln C(w,y)}{\partial \ln w_k} . \frac{C(w,y)}{w_k} \\ &= \left(\beta_k + \sum_l^n \gamma_{kl} \ln w_l + \sum_i^m \rho_{ki} \ln y_i \right) . \frac{C}{w_k} \end{aligned} \tag{2.7}$$

Optimal cost shares s_k of each input are given by the bracketed term, equal to $\frac{\partial\, ln\, C(w,y)}{\partial\, ln\, w_k}$. These cost shares describe the cost-minimising share of total expenditure each input should take up. Where the cost-minimising quantities and optimal cost shares are positive, costs increase with input prices.

Each hospital has different optimal cost shares dependent on output and wage levels. When these cost shares are evaluated, 6% of junior doctor cost shares are negative, 12% of nurses, and 53% of capital prices. Many observations do not have costs that increase with input prices. These observed departures from economic theory are likely due to two factors.

Firstly, the model takes capital and wage costs as price inputs but does not take material prices. Capital prices are derived from published financial statements and therefore adopt a narrow measure of capital as defined by accounting practice rather than a broader economic definition that would include working capital or other intermediary goods used in the production of healthcare. The omission of these inputs may create local departures from convexity.

Secondly, wage rate variations may reflect productivity differences within staff groups. Hospitals with higher wage rates have more productive staff at greater levels of seniority rather than higher wage rates due to local market conditions. As national remuneration frameworks cover all types of staff, differences observed may be attributable to differences in seniority and grade. Variation in wage rates may be partially the result of heterogeneous labour with different productivity levels, affecting the estimation results.

3. **Cost should be continuous and concave in input prices** — A function is continuous and concave if the Hessian matrix is negative semidefinite. Elements of the Hessian matrix H_{km} are given by:

$$\begin{aligned}
H_{km} =& \frac{\partial^2 C(w,y)}{\partial w_k \partial w_m} \\
=& \frac{\partial \left((\beta_k + \sum_l^n \gamma_{kl} \ln w_l + \sum_i^m \rho_{ki} \ln y_i) . \frac{C}{w_k} \right)}{\partial w_m} \\
=& \frac{\gamma_{km}}{w_m} . \frac{C}{w_k} + \left(\beta_k + \sum_l^n \gamma_{kl} \ln w_l + \sum_i^m \rho_{ki} \ln y_i \right) . \frac{1}{w_k} . \partial C / \partial w_m \\
& - \Delta_{km} \left(\beta_k + \sum_l^n \gamma_{kl} \ln w_l + \sum_i^m \rho_{ki} \ln y_i \right) . \frac{C}{w_k^2}
\end{aligned}$$

Where $\Delta_{km} =$

$$\begin{cases} 0 \text{ if } k \neq m, \\ 1 \text{ if } k = m. \end{cases} \tag{2.8}$$

is the Kronecker delta function. Substituting in the derived input demand function from Equation (2.7) yields:

$$\begin{aligned}
H_{km} &= \gamma_{km} \cdot \frac{C}{w_k w_m} + x_k . \frac{w_k}{C} . \frac{1}{w_k} . x_m - \Delta_{km} . x_k . \frac{w_k}{C} . \frac{C}{w_k^2} \\
&= \gamma_{km} \cdot \frac{C}{w_k w_m} + \frac{x_k \, x_m}{C} - \Delta_{km} . \frac{x_k}{w_k}
\end{aligned} \tag{2.9}$$

The Hessian matrix will be negative semidefinite only where the diagonal elements are negative. The main diagonal elements correspond to cases where $k = m$. 798 / 798 observations have negative elements for training grade doctors. However, only 48 / 798 capital observations and 0 / 798 nursing staff observations have negative diagonal elements.

4. **Cost should be non-decreasing in any output** — This condition is evaluated by looking at the individual cost-output elasticities calculated for each output category. Where these elasticities are negative, a percentage increase change in output causes a percentage decrease in cost. There are 6 output categories and 798 observations, giving 4788 total cost-output elasticities. Negative cost-output elasticities are present in 49 / 798 A&E observations, 7 /

798 critical care observations, 55 / 798 diagnostic observations, 0 / 798 elective inpatient observations, 0 / 798 non-elective inpatient observations, and 0 / 798 outpatient appointment observations. Apart from A&E and diagnostic services, there are few cases of non-compliance with this desired characteristic. Cost-output elasticities for A&E and diagnostic services are more likely to be negative than other outputs at particular observations as these elasticities are generally lower than other outputs. Their low level may be either because their outputs form a lower share of total costs or because they have low marginal costs. Consequently, local violations of cost function convexity are more likely to occur in these outputs, and these results are less concerning.

5. **Cost should be linearly homogeneous in input prices** — As the sum of coefficients on the input prices is less than 1, proportionate increases in all input prices do not raise the total cost by the same amount. This observation suggests that relevant input prices are missing, for example, consumables such as medicines. Assuming that each hospital faces similar prices in these areas from national procurement efforts, this should not unduly affect the estimation.

6. **No fixed costs in the long run** — Where the output variables are zero, the model's intercept term and other non-zero variables will ensure that costs are not zero. This condition can be tested asymptotically by considering the limit of the right hand side of the translog equation as the $\sum y_i$ approaches zero:

$$\begin{aligned}
&\lim_{\sum y_i \to 0+} (\alpha_0 + \sum_i^m \alpha_i \ln\, y_i + \sum_k^n \beta_k \ln\, w_k + \frac{1}{2}\sum_i^m \sum_j^m \delta_{i,j} \ln\, y_i \ln\, y_j \\
&\qquad + \frac{1}{2}\sum_k^n \sum_l^n \gamma_{k,l} ln\, w_k ln\, w_l + \sum_i^m \sum_k^n \rho_{i,k} \ln\, y_i ln\, w_k) \\
&= \lim_{\sum y_i \to 0+} \left(\sum_i^m \alpha_i \ln\, y_i + \frac{1}{2}\sum_i^m \sum_j^m \delta_{i,j} \ln\, y_i \ln\, y_j + \sum_i^m \sum_j^n \rho_{i,j} \ln\, y_i ln\, w_i \right)
\end{aligned} \tag{2.10}$$

The limit of ln 0 from above being $-\infty$, this becomes

$$\begin{aligned}
&= -\infty \sum_i^m \alpha_i + \infty \frac{1}{2} \sum_i^m \sum_j^m \delta_{i,j} - \infty \sum_i^m \sum_k^n \rho_{i,k}\, ln\, w_k \\
&= -\infty \left(\sum_i^m \alpha_i - \frac{1}{2} \sum_i^m \sum_j^m \delta_{i,j} + \sum_i^m \sum_j^n \rho_{i,j}\, ln\, w_i \right)
\end{aligned} \tag{2.11}$$

The predicted cost from the multiproduct translog consequently asymptotically approaches zero only where the bracketed term is positive, which leads to ln $C(w, y) =$

$-\infty$ and $C(w, y) = 0$. The estimated translog cost function accordingly does not meet this criterion.

Overall, the estimated cost function performs better with anticipated properties relative to output variables than price variables. However, to estimate scale economies, it is more important that the function behaves better in the output spaces. Due to their national determination, there is limited variation in wage rates in the English case. Consequently, there is unlikely to be a relationship between wages and scale that may be present in decentralised systems, for example, where hospitals have a local monopsony on certain staff groups. The observed departures from wage convexity and demand are, therefore, less salient for our purposes and are acceptable given the flexibility of the translog form estimated.

2.5.3 Visualisation of Modelled Costs

Figures 2.6 - 2.10 show a close relationship between predicted and actual costs. This close relationship stems from the number of terms in the model, so we should be cautious about overfitting, given the high R^2 statistic and the large intercept term. The close relationship may impair the applicability of the function to output ranges outside of those observed. However, looking at the ability of the estimated function to model the observed costs, a close relationship can be observed in each year and across each output type, covering all reasonable output ranges.

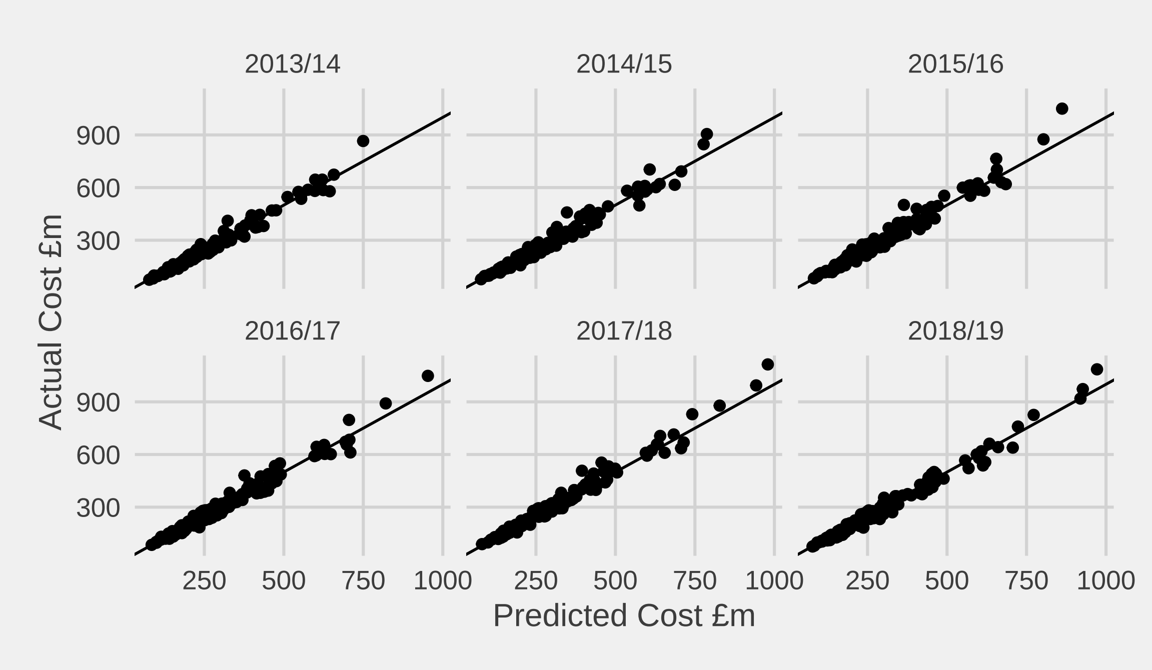

Figure 2.6: Predicted Vs Actual Cost in each Year (Translog). Predicted costs are closely correlated to actual values. Predictions are derived from the estimated cost function and actual output levels

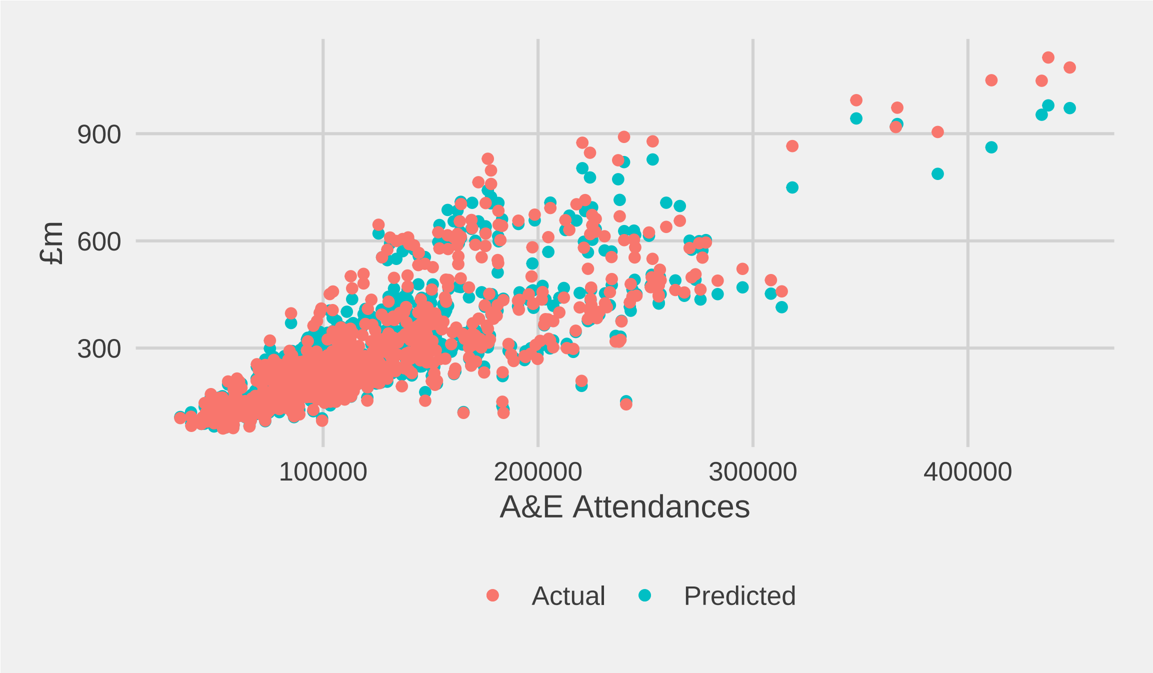

Figure 2.7: Actual and Predicted Costs vs Activity - A&E Attendances

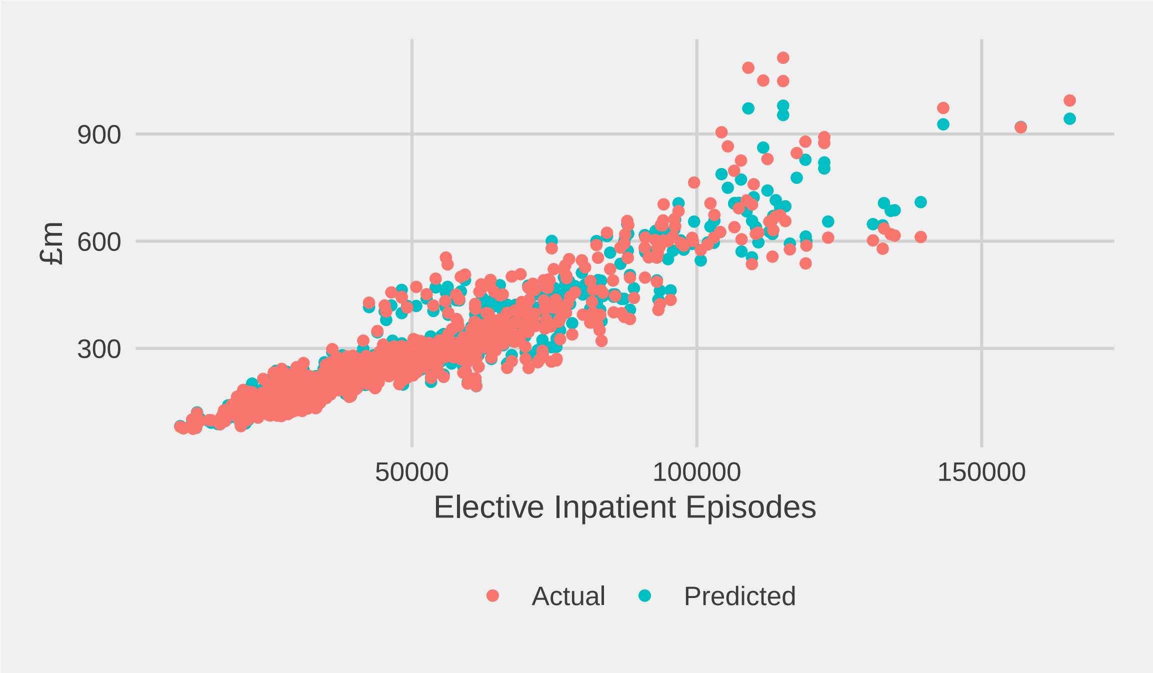

Figure 2.8: Actual and Predicted Costs vs Activity - Elective Inpatient Episodes

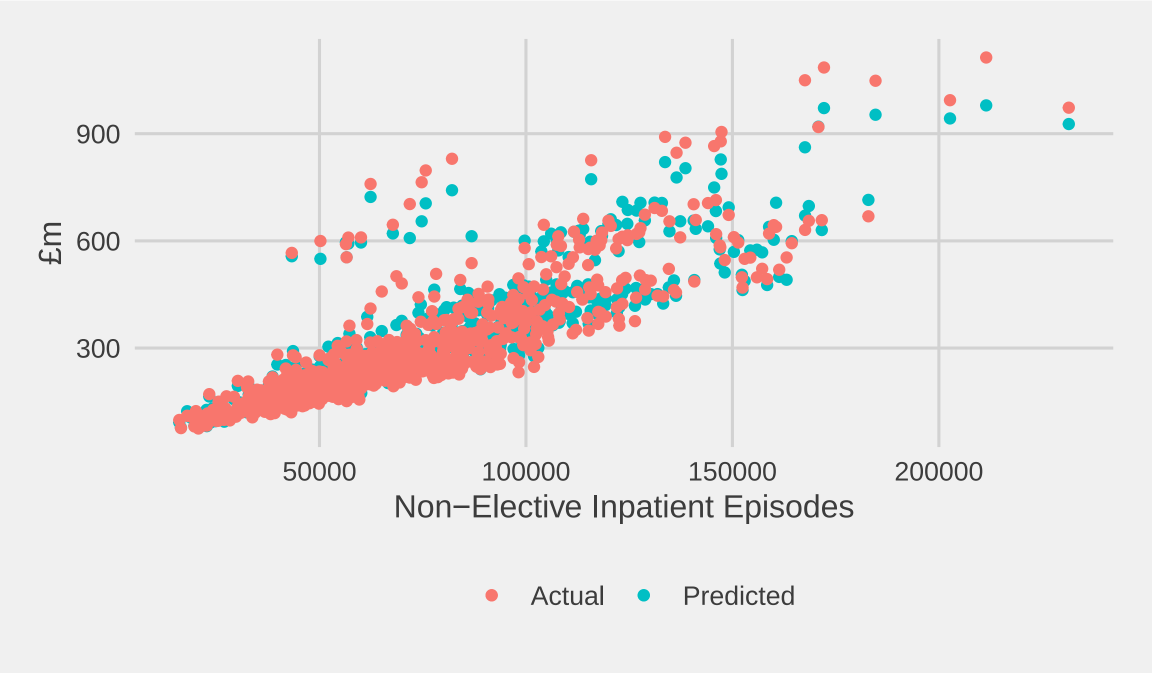

Figure 2.9: Actual and Predicted Costs vs Activity - Non-Elective Inpatient Episodes

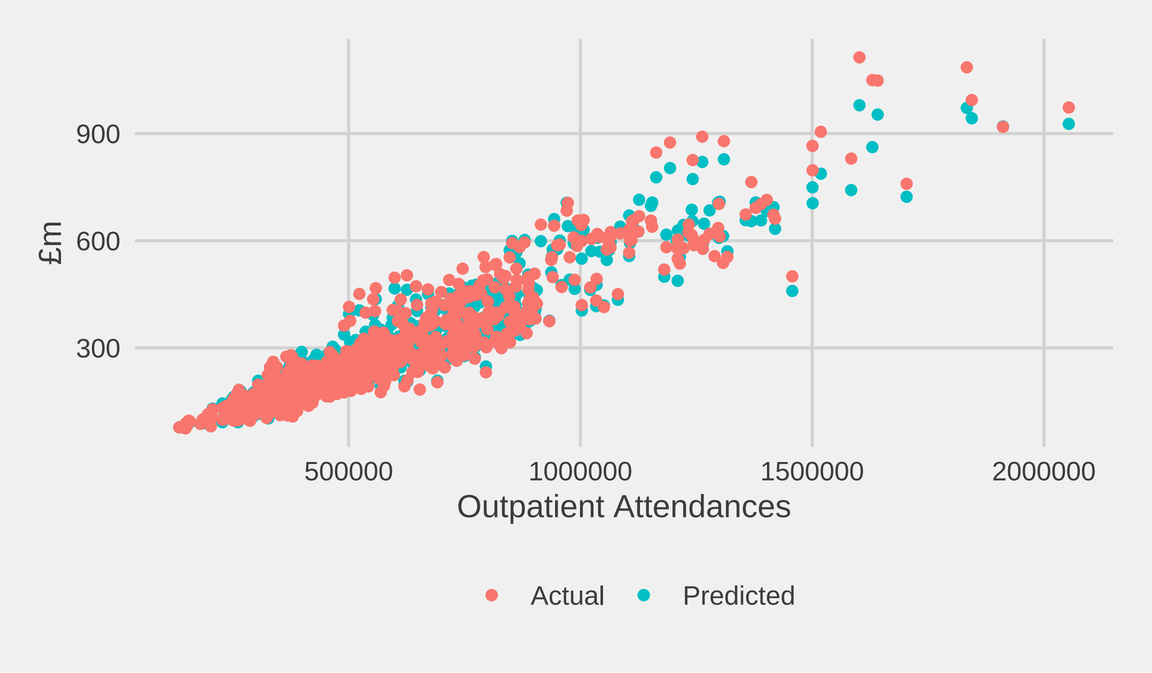

Figure 2.10: Actual and Predicted Costs vs Activity - Outpatient Attendances

2.5.4 Average and Marginal Costs

As noted in Section 1.3, it is difficult to define average and marginal costs for multi-product cost functions. Figures 2.11 - 2.16 show a version of firm-specific marginal costs for each output, varying one of the output categories by increments of one output unit whilst holding all other output categories constant at observed levels. Predicted total costs for each output level are subtracted from the predicted costs where that output is one unit lower, giving the marginal cost. In the plots, 30 randomly selected organisation-year observations are selected to examine the shape of the marginal cost curves around observed output levels. Individual observations are differentiated by colour, as each organisation-year observation has a different marginal cost curve dependent on production levels of other outputs and the input prices they face. Average costs are defined similarly, altering one output and computing changes to predicted total costs divided by the cumulative departure from the observed output. This method gives a ‘ray’ average cost defined by the expansion of a plane from the current output along the chosen output axis. These definitions can be used to examine how the cost curves change with variation in output levels.

Elective inpatient average and marginal costs are generally flat as output varies, with minimal variation noted around current output levels. Non-elective average and marginal costs decline with output, whilst outpatient costs tend to increase. The greater decline of non-elective costs with output may reflect the nature of non-elective care. Hospitals must maintain a greater capacity for non-elective admissions than elective care, where small waiting lists can smooth out temporary demand and supply mismatches and minimise costs (Siciliani et al., 2009). In addition, there are fewer opportunities for amalgamating non-elective care services where travel time to the hospital is important. The flat nature of outpatient costs reflects the nature of the healthcare product supplied, typically a short, set interval of time in consultation with a doctor. Consequently, the cost base is relatively simple and invariant with output volume.

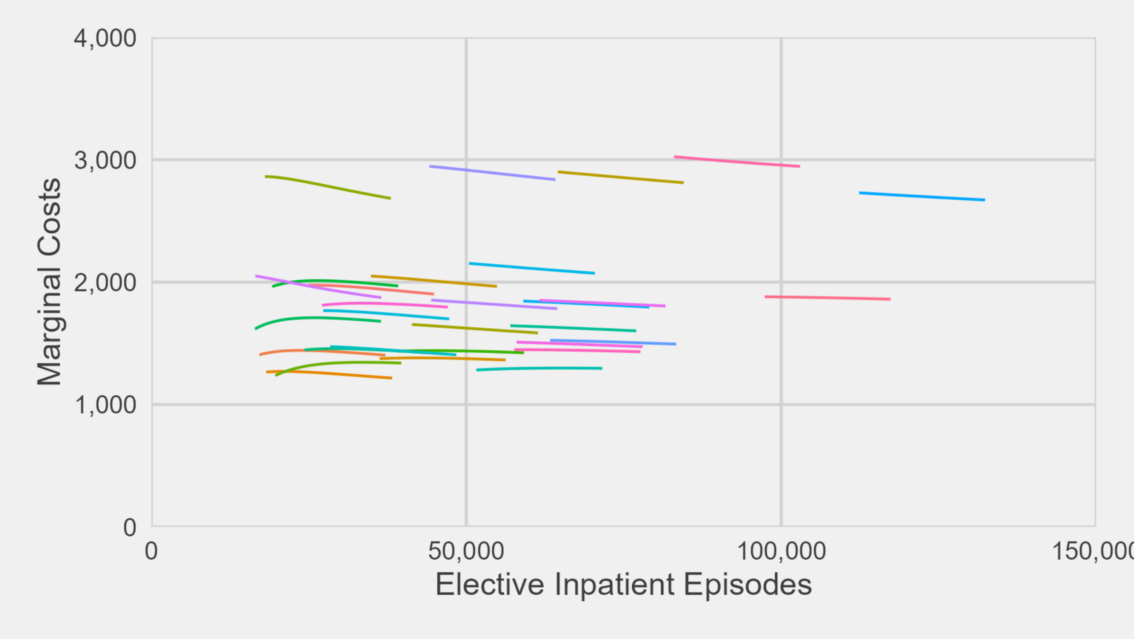

Figure 2.11: Elective Inpatient Episode Marginal Cost Curves (Translog). Marginal cost curves are derived from the estimated cost function by calculating the marginal cost effects of incremental changes in one output type, holding other outputs constant. The plot shows sections of cost curves generated from a random sample of 30 observations, differentiated by colour

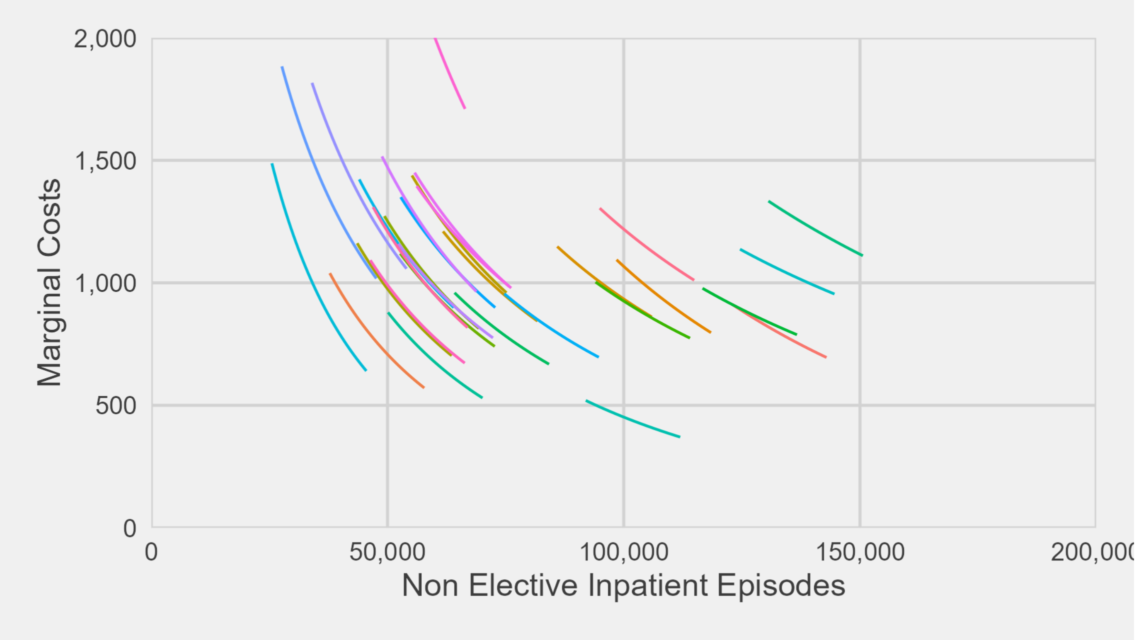

Figure 2.12: Non-Elective Inpatient Marginal Cost Curves (Translog). See Fig 2.11 for details on derivation.

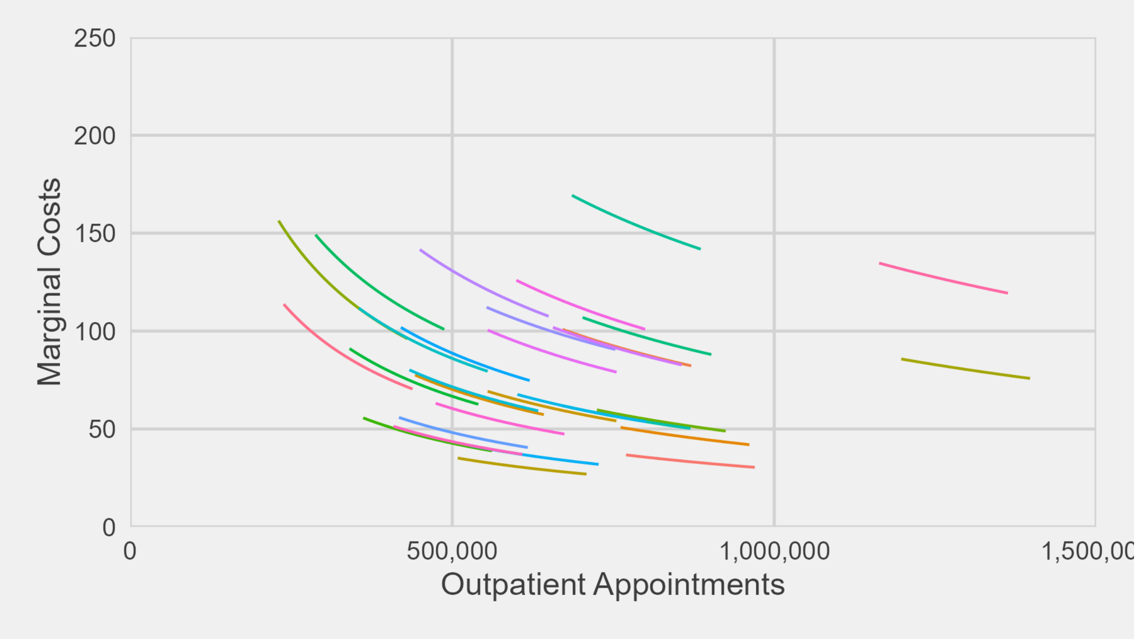

Figure 2.13: Outpatient Marginal Cost Curves (Translog). See Fig 2.11 for details on derivation. Outpatient appointment costs are relatively invariant to output levels. Appointments are a discrete period of time in consultation with a doctor with consequent limitations on scale economies.

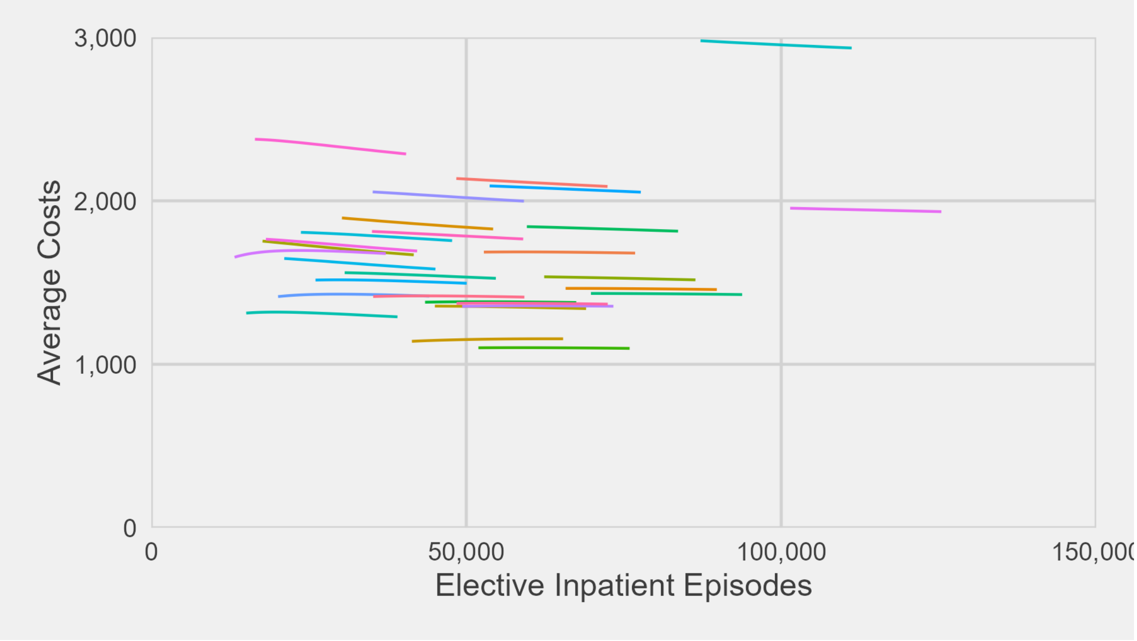

Figure 2.14: Elective Inpatient Average Cost Curves (Translog). Average cost curves are derived from the estimated cost function by calculating the total cost effects of discrete changes in one type of output and then dividing by the change in output, holding other outputs constant. The plot shows sections of cost curves generated from a random sample of 30 observations, differentiated by colour

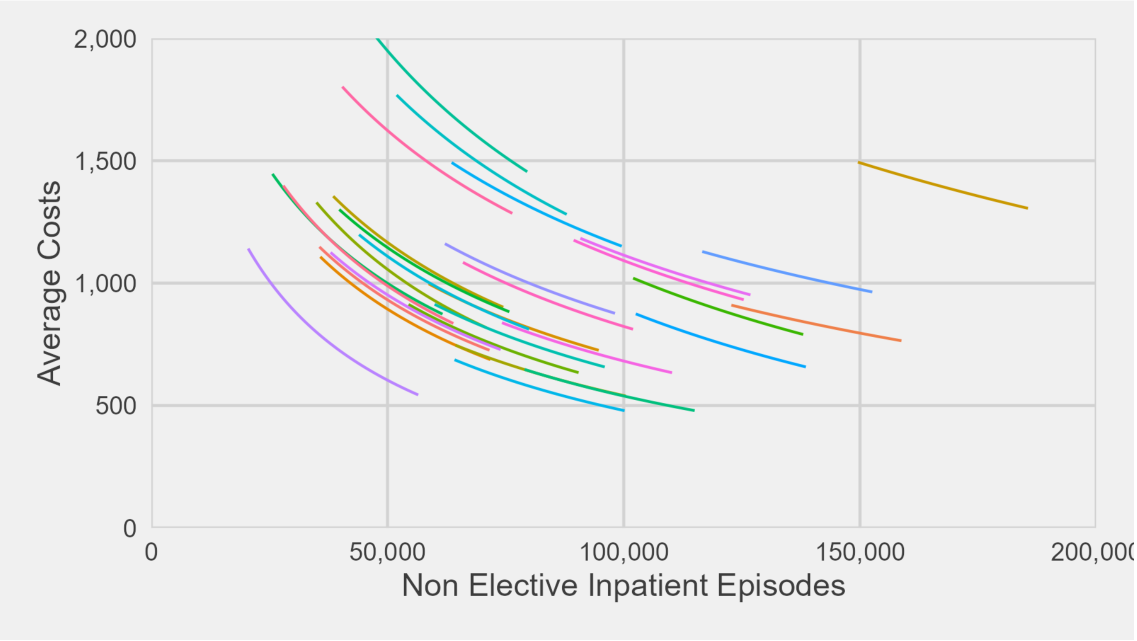

Figure 2.15: Non-Elective Inpatient Average Cost Curves (Translog). See Fig 2.14 for details on derivation.

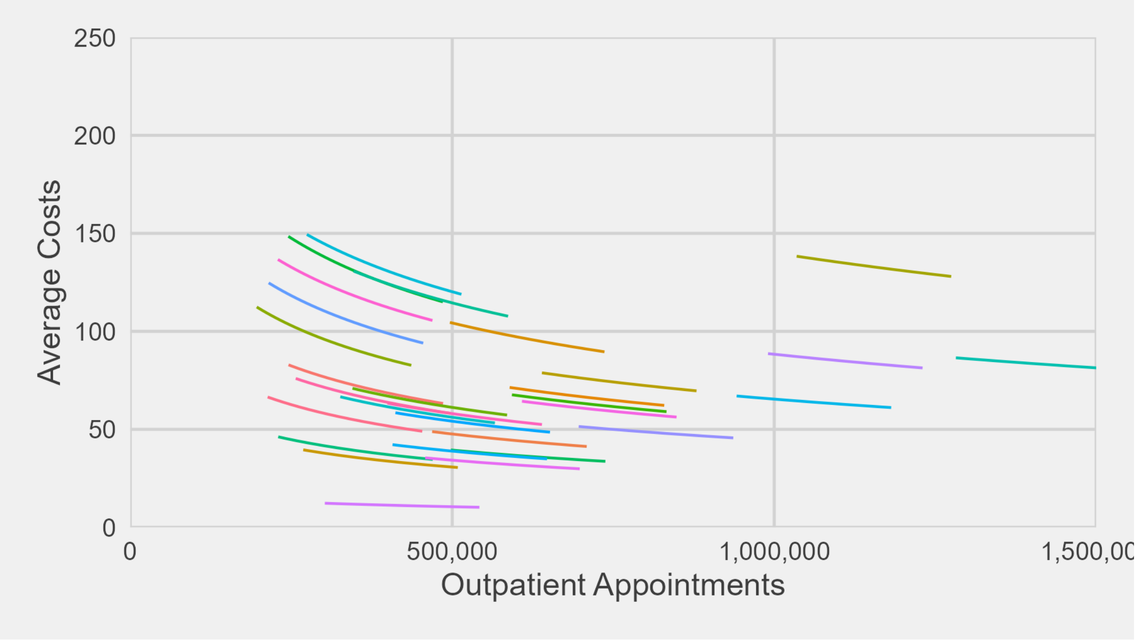

Figure 2.16: Outpatient Average Cost Curves (Translog). See Figure 2.14 for details on derivation. Outpatient appointment costs are relatively invariant to output levels. Appointments are a discrete period of time in consultation with a doctor with consequent limitations on scale economies.

2.5.5 Scale Economies

The scale economy index is computed according to Equation (2.6), using the coefficient estimates in Table 2.7 and actual output and input prices observed for each observation in the dataset. Standard errors for the scale economy index are derived via the delta method as a combination of the coefficient estimates.

Figure 2.17 shows that the distribution of scale economies falls close to constant scale economies, though more organisations lie above the line and have positive economies of scale. The median scale elasticity is 1.04, so a proportional 1.04% expansion of output for the median hospital would incur a 1% increase in cost.

A tabulation of average scale economies is set out in Table 2.8, showing the mean average of the scale economy measure for each organisation over the years in the dataset. 50 organisations have an average scale economy measure indicating economies of scale. An organisation is defined as having economies of scale where the low point of the 95% confidence interval is above 1 in the majority of observations. Analogously, an organisation is defined as having diseconomies of scale where the high point of the 95% confidence interval is below 1 for the majority of available observations. Where the 95% confidence interval spans 1 in the majority of observations for that organisation, it is categorised as operating under constant scale economies. According to this definition, 90 organisations have approximately constant scale economies, 1 organisation had a mean high point of the 95% scale economy measure below 1 and therefore experienced diseconomies of scale, and 50 organisations operated at a level of output with economies of scale.

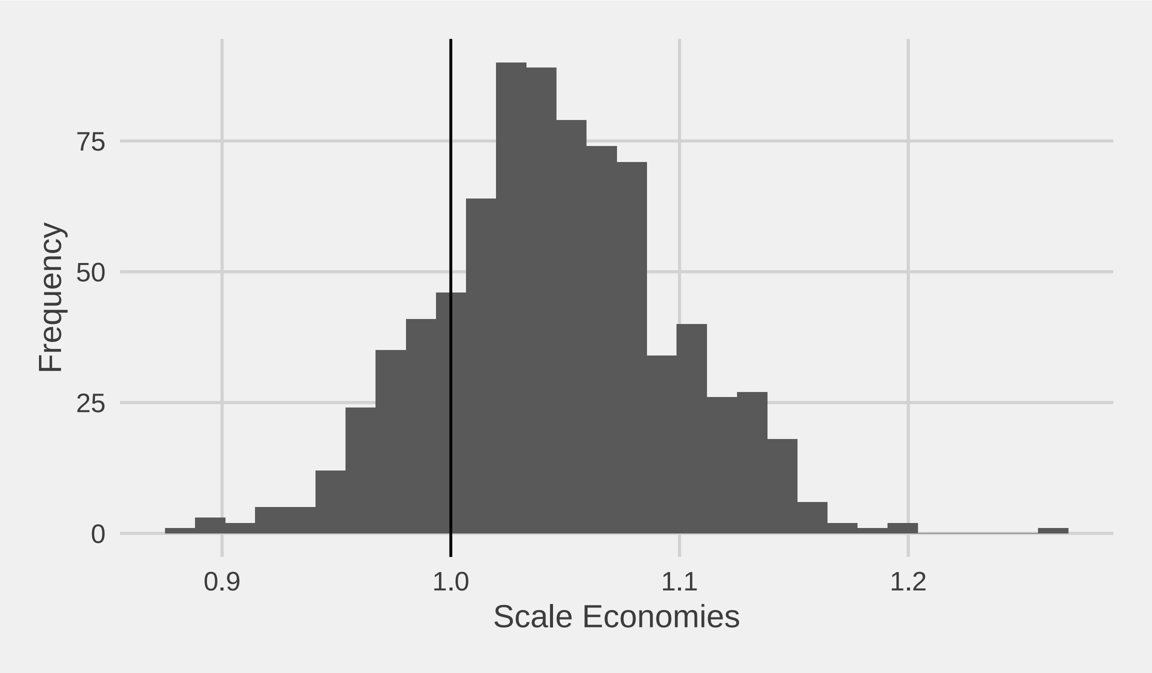

Figure 2.17: Distribution of Scale Economy Index (Translog). Most organisations have small but positive estimated scale economies

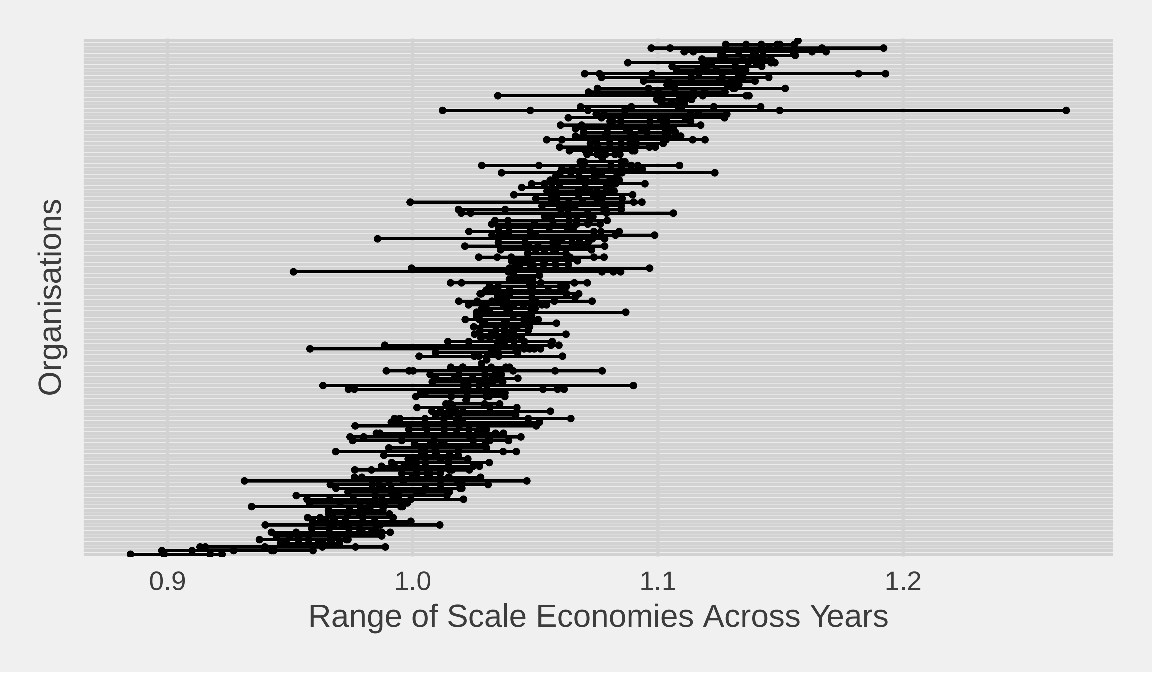

Figure 2.18: Range of Scale Economy Indices by Organisation across Years (Translog). Scale Economies are relatively stable for each organisation across the period

Table 2.8: Average Scale Economies (2013/14 to 2018/19)

Average Scale Economies	**N**	**Percentage**
Constant economies of scale	90	64%
Diseconomies of scale	1	1%
Economies of scale	50	35%
Total	**141**	**100%**

Observed scale economies are plotted against outputs and other size measures to investigate the relationship between size and scale economy. Measures of scale other than outputs, such as bed numbers, are frequently used in the literature as they avoid the "regression fallacy" identified by Friedman (1955). This issue, and the wider validity of proxy capital measures, is discussed in more detail in Section 1.4.7. Scale economy results are presented against both output and beds below. This approach has the advantage of offering multiple criteria to judge expected scale economies. Where no definitive measure of size exists, using multiple measures alleviates biases with particular measures of size.

Figure 2.19 visualises the relationship between scale economies and organisational size defined by bed numbers. A polynomial of degree 2 fitted to the data in figure 2.19 crosses the boundary between positive returns to size and negative returns at approximately 1,200 beds, though there is variation amongst organisations. Uncertainty in the estimates is quantified by plotting 95% confidence intervals around each data point based on the linear combination of parameter values evaluated at actual output levels. Each observation has a large uncertainty interval, but there are a consistent number of observations above the line of constant scale economies before the minimum efficient scale is reached.

Using fixed asset valuations as an alternative measure of size returned similar results, as bed numbers and fixed asset values are strongly correlated (Pearson's r - 0.729). Figures 2.20 and 2.21 plot scale economies against output levels for elective inpatient episodes and outpatient attendances. Polynomial curves are fitted to the point scale economy estimates. Overall scale economies move from increasing to decreasing at approximately 90,000 elective inpatient episodes and 1,000,000 outpatient attendances.

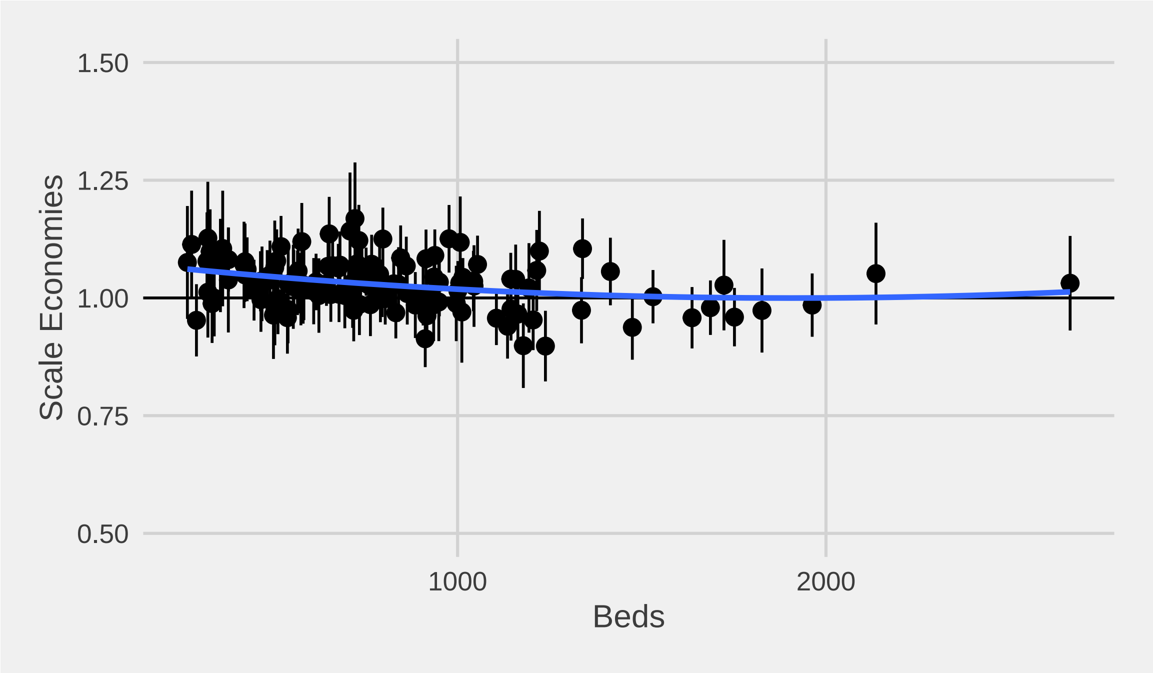

Figure 2.19: Scale Economy Index vs Bed Numbers (Translog). 2018/19 observations only

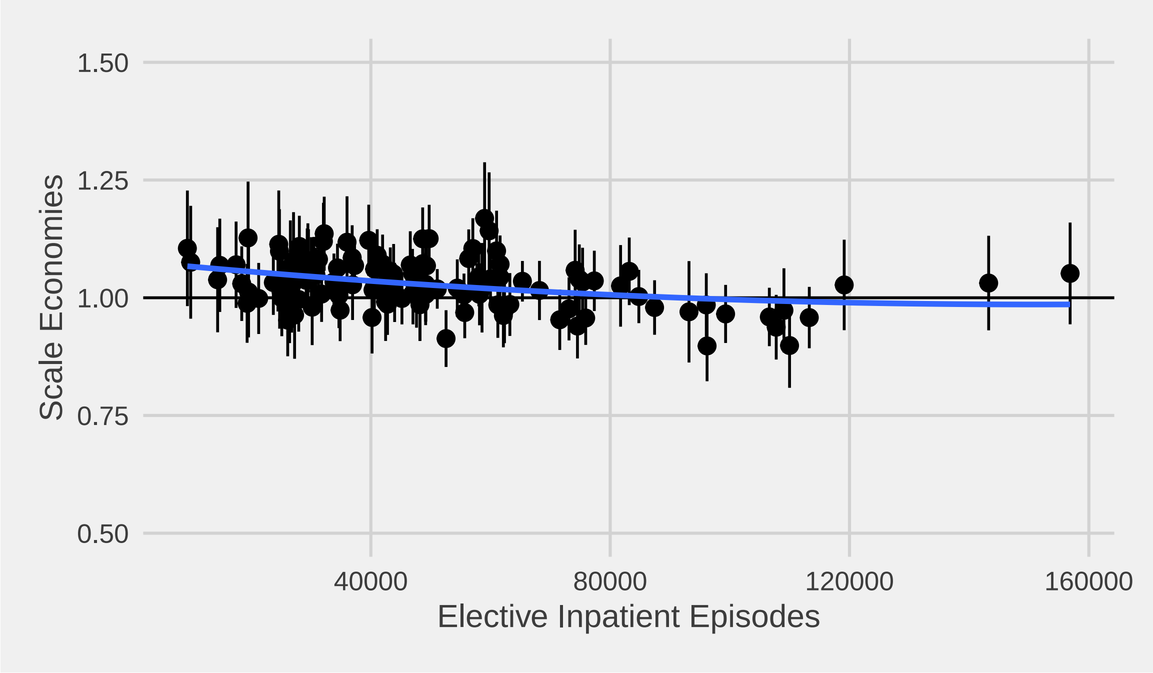

Figure 2.20: Scale Economies vs Inpatient Episodes (Translog). 2018/19 observations only

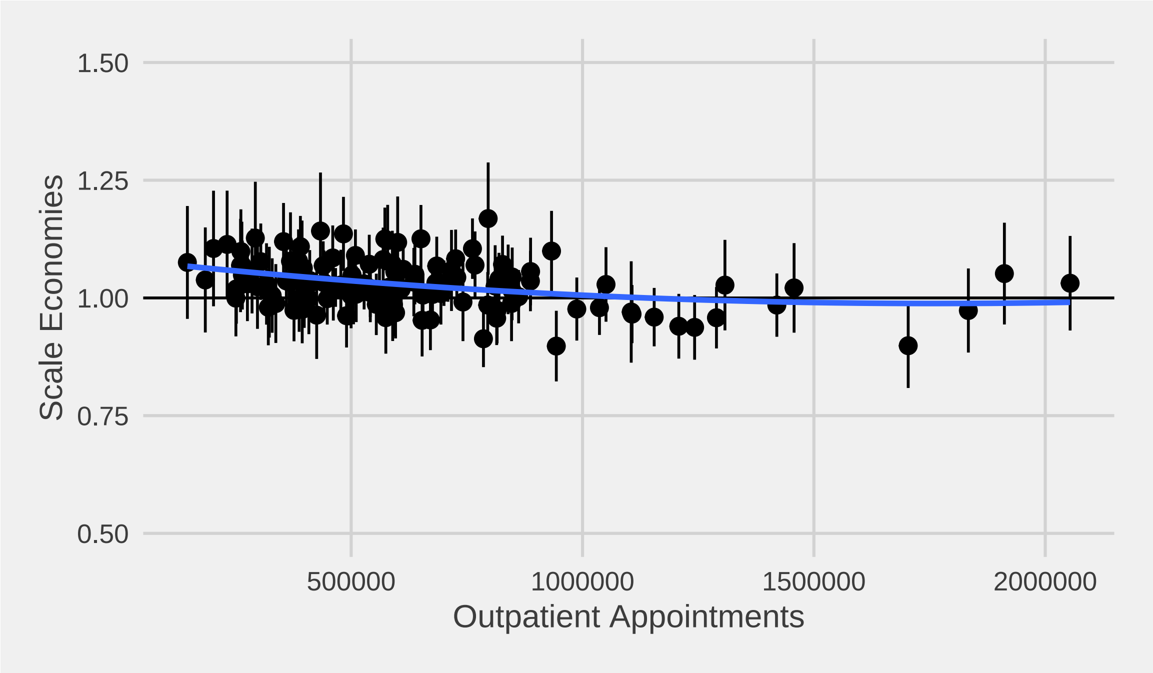

Figure 2.21: Scale Economies vs Outpatient Attendances (Translog). 2018/19 observations only

Disaggregating the graphs by year also shows that the general result is consistent across the years in the dataset. Figure 2.22 shows scale economies against average beds in each dataset year, with similar results across years. This observation suggests that technological change and relative prices were stable over the period and that averaging scale economies over the years is reasonable and not likely to be affected by any particular years with large outliers.

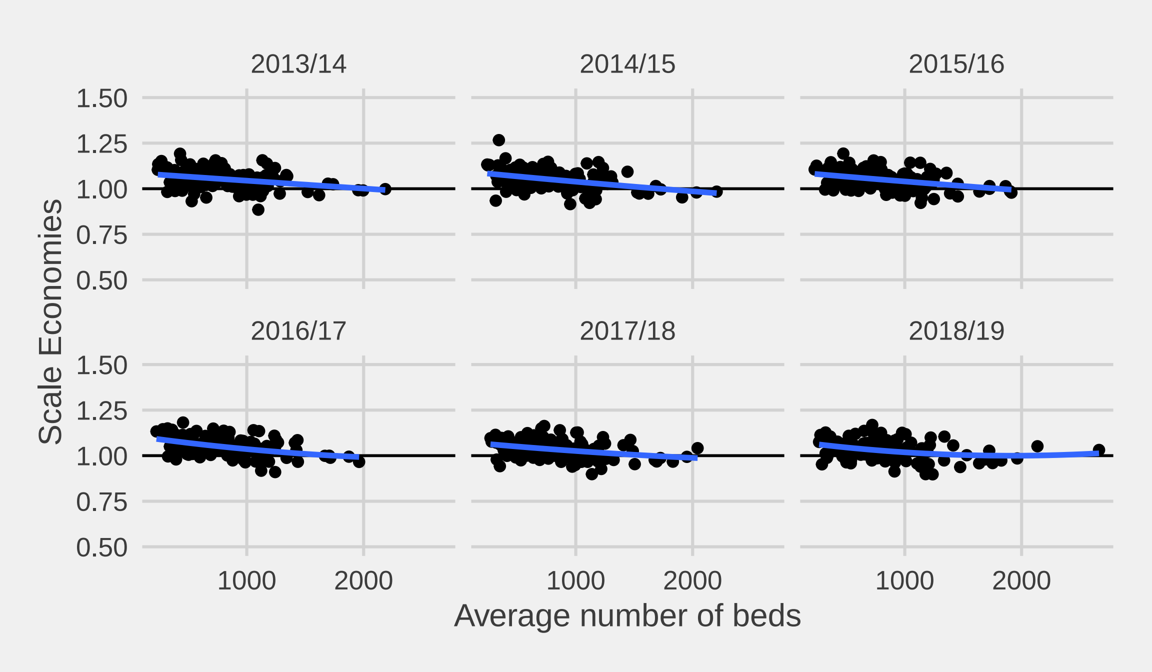

Figure 2.22: Scale Economies vs Bed numbers in each Year (Translog): The relationship is relatively consistent across each year in the panel

Scale economies are observed to be lower for organisations close to the high-cost area of London, where labour costs attract a premium due to a higher local price level and higher labour demand. Scale economies are lower irrespective of the size of the organisation. Figure 2.23 illustrates this by plotting a version of Figure 2.19, showing scale economies against beds with regions differentiated by colour. London scale economies are consistently below other hospitals as they have higher labour input prices. These higher labour prices mean that proportionate increases in output raise total costs more than other hospitals.

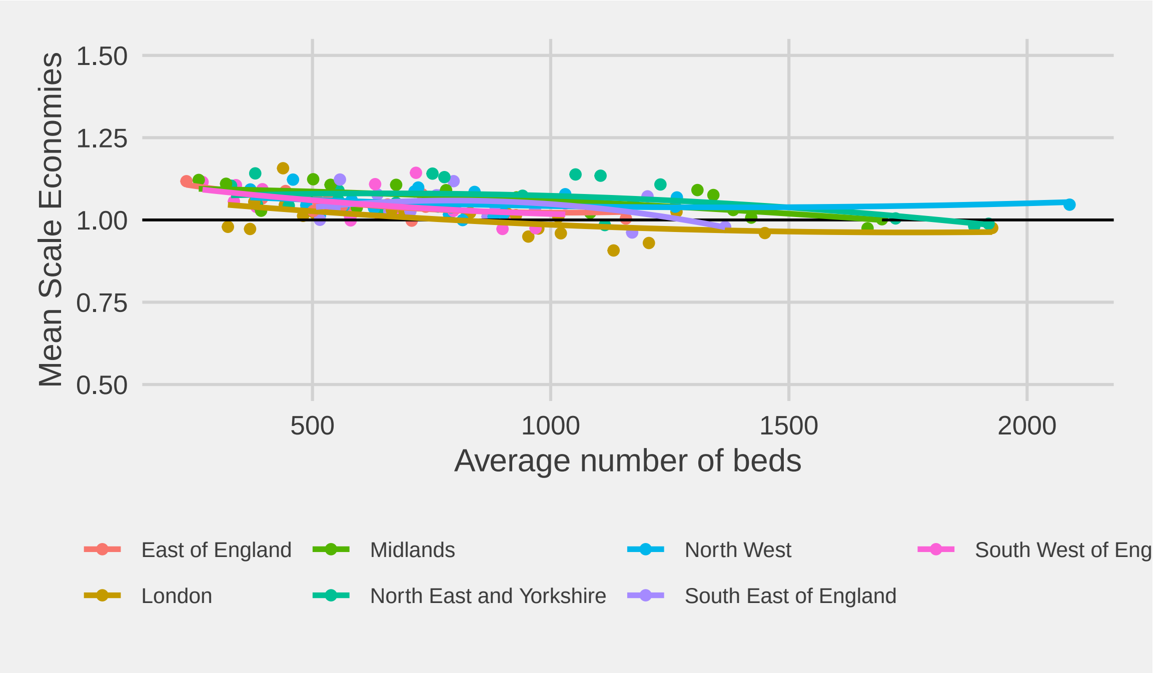

Figure 2.23: Scale Economies by NHS England Region (Translog). Organisations in London and neighbouring areas in the South East are required to pay higher wages in the form of a wage supplement for high-cost areas. This supplement is intended to compensate staff required to work in higher-cost areas, effectively raising variable costs for organisations in London or proximate to it

2.6 Discussion

This chapter estimates scale economies for English NHS hospitals using a multi-product translog cost function. Most organisations are operating either at constant or positive scale economies. The results suggest that the point of diminishing scale economies starts around 1,200 beds, or using other measures of scale, 90,000 admitted elective episodes or 1,000,000 outpatient episodes. This limit is in excess of the median hospital, which had 740 beds in 2018/19. Limits to scale economies are higher than previous estimates in the literature, perhaps attributable to technological developments since previous studies. Economies of scale were also lower in higher cost areas (London) due to higher wage rates increasing costs to a greater extent for comparable increases in output.

The obtained results update older estimations of hospital scale economies and provide evidence from a country not often studied. This study also improves on previous estimates by using more detailed output and cost information and including input prices. A fuller range of hospital outputs is considered, including admitted and ambulatory care. Output homogeneity is dealt with by a more detailed categorisation of outputs and using HRGs to re-weight outputs according to expected resource intensity.

A possible limitation of this study arises from the lack of a control variable for hospital quality, which could be a possible source of endogeneity if hospital quality is associated with healthcare costs (in either direction). If hospitals supplying better quality care can attract more patients, then size and care quality will be correlated. Should healthcare quality and costs be related, then there is a potential source of bias from the omission of a variable measuring care quality. The relationship between quality and referral patterns is particularly complex, however. There are many reasons why patients and their referring primary care doctors may choose a particular hospital. Studies have found that the strongest effect on patient choice is distance (Smith et al., 2018), though waiting times and local competition are also explanatory factors, as well as quality (Moscelli et al., 2016). Choosing an adequate quality measure is also difficult at an aggregated hospital level, especially where unobserved factors such as the relative sickness of patients in the area can affect quality measures such as mortality or readmission rates. Patients (or their referring doctors) may be more influenced by condition-specific indicators that are more detailed than overall measures of quality (Gutacker et al., 2016). Finally, when the relationship between cost and quality is empirically tested, results are unclear or non-linear (Gutacker et al., 2013), though some studies suggest that higher quality care is associated with higher costs (Carey & Stefos, 2011). Given these difficulties with finding a usable measure of quality, the inclusion of a quality variable is omitted in the analysis.

A further limitation of the approach followed is that the estimation of a long-run cost function assumes that organisations currently cost-minimise and have capital levels that are optimal for their current output. Other authors have criticised the approach of directly estimating long-run cost functions because the assumption is not necessarily true (Aletras, 1999). Capital levels may be slow to respond to short-run output changes, meaning that observed values used in the estimation may be higher than the long-run optimal, where hospitals can choose the optimal capital level.

Policymakers could look to realise some of these gains by consolidating smaller and mid-size organisations into larger entities or removing services from organisations where diseconomies of scale apply. Differences between the high-cost area of London and the rest of the country suggest that expansion outside of London is likely to yield greater benefits, though there are obvious geographic limits to how far patients would be able or willing to travel. Average costs are observed to decline faster with size in the case of non-elective cases, where proximity to immediate treatment is necessary, suggesting that attempting to achieve benefits from increasing scale may be further limited by the need to ensure access to hospital healthcare.

Future research could generalise the approach to estimate short-run cost functions by holding capital (proxied by beds or financial statement asset values) constant, deriving optimal short-run capital levels and then adopting an envelope condition to estimate the long-run case. Other potential future research includes contrasting translog results with those obtained by a quadratic cost function specification, non-parametric methods such as data envelopment analysis, and estimating scope economies using this dataset.

Chapter 3

Economies of Scale in English Hospitals: A Quadratic Estimation

3.1 Introduction

Two main functional forms are used in the literature for estimating scale economies. Chapter 2 uses the more common translog form, which is used by most studies estimating hospital scale economies. The typical alternative is the quadratic functional form, particularly used in studies seeking to indirectly estimate long-run cost functions via an envelope of short-run functions.

This chapter uses the same dataset and methodology from Chapter 2 to estimate a quadratic rather than a translog cost function. Some *a priori* reasons to prefer the quadratic or translog functional form are discussed, for example, the case of data containing zero values which cannot be defined when log-transformed. The results from both translog and quadratic estimations are compared against the desirable theoretical properties of a cost function, followed by a discussion of the regression results generally and an interpretation of marginal effects implied by these results. Finally, scale economies are computed using the alternate quadratic specification and investigate the relationship of scale economies with size, as measured by the number of hospital beds.

The results show that both specifications fit the data in a comparable way and return similar scale economy values overall. There are some differences in the relationship between scale economies and size, with the quadratic function showing lower scale economies than the translog amongst smaller trusts and higher scale economies amongst larger trusts. Consequently, whilst the translog specification showed scale economies declining with size, the quadratic form shows that more hospitals currently operate under constant returns to scale, with no apparent limit to scale economies as size increases.

3.2 Previous Literature

Many studies have used translog functions to measure hospital costs (Azevedo & Mateus, 2014; Conrad & Strauss, 1983; Cowing & Holtmann, 1983; Scott & Parkin, 1995). The alternative approach is the quadratic functional form, particularly used in more recent studies seeking to indirectly estimate long-run cost functions (Kristensen et al., 2012; Preyra & Pink, 2006). Like the translog form, the quadratic cost function allows average costs to vary non-linearly with outputs, so estimates for economies of scale can also vary with size. To the best of my knowledge, there are no empirical studies which contrast results of two different functional forms.

Re-estimation using identical data but differing cost function forms is helpful for two reasons. Firstly, it offers a robustness check for the main result in Chapter 2 in case the observed results are highly variable with a different specification. If results were observed, it would suggest that uncertainty in the main Chapter 2 result could be high. Giannakas et al. (2003) found that the choice of form can affect results for efficiency calculations, so scale economy calculations could also be similarly affected. Secondly, it is useful in and of itself to investigate how changing the form could affect results, so we can better understand the effects of using different cost function forms and how studies that use only one method could be affected by the form choice.

3.3 Data

The data used in this chapter is identical to that of Chapter 2. This chapter considers an alternate form of the cost function only.

3.4 Method

As set out in Chapter 1, the quadratic cost function is one of two commonly used forms which allow scale economies to vary with size. The quadratic cost function is similar to the translog form, specifying total cost as a function of outputs, input prices, and cross-product terms. The quadratic principally differs from the translog in the absence of logged variables on either side of the cost function.

To illustrate, the multiproduct translog form used in Chapter 2 is:

$$\begin{aligned}\ln\, C_{ht}(y,w) =&\alpha_0 + \sum_i^m \alpha_i \ln\, y_{iht} + \sum_k^n \beta_k \ln\, w_{kht} + \frac{1}{2}\sum_i^m\sum_j^m \delta_{ij} \ln\, y_{iht} \ln\, y_{jht} \\ &+ \frac{1}{2}\sum_k^n\sum_l^n \gamma_{kl}\, ln\, w_{kht}\, ln\, w_{lht} + \sum_i^m\sum_k^n \rho_{ik} \ln\, y_{iht}\, ln\, w_{kht} + \epsilon_{ht}\end{aligned} \tag{3.1}$$

Where y_i is the amount of the ith output produced (of m total outputs), w_i is the ith input price (of n inputs), and α_i, β_k, δ_{ij}, $\gamma_{k,l}$, and $\rho_{i,k}$ represent coefficients estimated in the model. Hospital index h and time period t reflect the panel data nature of the dataset. The error term $\epsilon_{ht} = \tau_h + v_{ht}$ comprises a term τ_h representing hospital specific effects and a random noise term v_{it}.

In this chapter, the analogous quadratic form is used, identical other than the lack of log transformation:

$$\begin{aligned} C_{ht}(y,w) =&\alpha_0 + \sum_i^m \alpha_i\, y_{iht} + \sum_k^n \beta_k\, w_{kht} + \frac{1}{2}\sum_i^m\sum_j^m \delta_{ij}\, y_{iht}\, y_{jht} \\ &+ \frac{1}{2}\sum_k^n\sum_l^n \gamma_{kl}\; w_{kht}\, w_{lht} + \sum_i^m\sum_k^n \rho_{ik}\, y_{iht}\, w_{kht} + \epsilon_{ht} \end{aligned} \tag{3.2}$$

For reasons of clarity, time and individual subscripts are omitted from this point on in the chapter.

As in Chapter 2, a dummy variable for teaching status is added to capture any residual teaching costs. To deal with issues of colinearity, output and input variables are centred, following the approach taken with the translog specification. The cost function is estimated using a random effects estimator and robust standard errors.

As noted in Chapter 1, the choice of functional form of the cost function has implications for cost function estimation. The translog and quadratic specifications have different properties that make them more or less desirable depending on the relative importance of each property. The logarithmic terms used in the translog allow a constrained optimisation which preserves linear homogeneity in input prices. This preservation is achieved by constraining the model such that the sum of coefficients on (log) input prices sum to one, so that total cost is homogeneous of degree one in prices, and a set percentage increase in wages and prices is reflected in the same percentage increase in total cost. This constraint is not possible in the linearised quadratic. However, the quadratic form allows easier estimation of short-run and long-run versions of the cost function, as the presence of logged values can complicate partial derivatives. The use of a quadratic form, where values are not logarithms, also allows us to model instances where certain outputs are zero. These zero-level outputs would be undefined in the translog due to the log terms.

At the level of aggregation used in this and preceding chapters, there are no zero output variables. However, were we to include the specialist trusts from Appendix A, there would be 12 hospitals with zero values for A&E Attendances, as they have no A&E department. 2 of the 12 also have no Critical Care activity, lacking Critical Care wards. If activity were aggregated into categories based on clinical specialty, as is the case in Chapter 5 then 2 hospitals have zero values for Obstetrics and Gynaecology. In this case, the quadratic needs to be used for the cost function to be defined.

In the results section, marginal cost derivatives are discussed, along with an evaluation of the estimated quadratic cost equation against the six desirable cost function properties in Chapter 1. Average and marginal cost expressions are derived, discussed and visualised. Scale economies are then computed from the cost function and contrasted with prior results using a translog functional form. Scale economies are derived using the same methodology as Chapter 2, except that the relevant formula changes for the quadratic form. Rather than being $\sum_i \frac{\partial\ C}{\partial\ y_i}$, in the case where terms are not logged, the size elasticity used to calculate scale economies is $\sum_i \left(\frac{\partial C}{\partial y_i} * \frac{y_i}{C}\right)$. Apart from this difference, methods to calculate scale economies are identical.

3.5 Results

3.5.1 Regression Results

Table 3.1 shows the estimation results. Centring the output and input variables means the intercept term £302,301,208 is comparable to the mean hospital total cost (£297,942,149). The coefficients on each output variable reflect marginal costs evaluated at the sample mean. Similarly, these costs are comparable to the reimbursement values for each activity in the case of A&E admissions, Critical Care episodes, elective and non-elective inpatients, and outpatient appointments. Not all diagnostic services are directly remunerated as they are trivially inexpensive, as reflected in the low coefficient in the regression. Though costs are low, the large volume of diagnostic tests mean that the contribution of this area is nontrivial, though not as large a contributor as admitted care and outpatients the larger areas (see Table 2.2). Coefficients on the doctor and nursing wage variables reflect marginal effects on the total cost of a £1 rise in wages for each labour type. These coefficients are comparable to the average number of staff covered by the relevant pay framework for these groups. The cost of capital coefficient was not statistically significant.

Amongst interaction terms, A&E and Critical Care Episodes have a negative coefficient, which was negative (as was the case in the translog specification - see Table 2.7). In the quadratic specification this is statistically significant. As A&E Attendances are likely to precede a critical care stay, those admitted to Critical Care via A&E may have a less costly A&E treatment than others, perhaps because those requiring critical care following an A&E attendance are unlikely to spend much time in A&E itself before admission, in comparison with lower acuity cases who may receive ambulatory care in A&E. The sign on the interaction between Critical Care Episodes and Non-Elective Episodes is positive, suggesting that those admitted non-electively do have more expensive Critical Care overall. The interaction of A&E with diagnostic services was also statistically significant and positive but was very low in magnitude. A&E patients, who may not have a diagnosis on presentation, are perhaps more likely to require expensive diagnostics.

The self-interaction term for Critical Care Episodes is negative, suggesting decreasing costs with scale, at least at the mean level of outputs. The sign on the interaction between Critical Care Episodes and Non-Elective Episodes is positive, suggesting that those admitted non-electively have more expensive critical care than elective cases. Critical Care and Outpatients also had negative spillovers, which is non-intuitive to explain but may be attributable to casemix, with those cases requiring critical care having an extended period of convalescence and post-discharge hospital care in an outpatient setting. Diagnostic services were associated with higher costs amongst Non-Elective Inpatient Episodes and also had a positive term on the self-cross-product. However, the magnitudes of these coefficients are very low, so these effects may not be influential except at high levels of diagnostic activity.

Higher nursing wages were associated with higher cost A&E attendances, higher costs in Elective Episodes, and lower costs in Non-Elective care. However, the magnitude of the effect in non-elective care is much lower than that for elective care. It may be that the relative share of nursing labour in output is higher in A&E and Elective Inpatient care than in other forms of activity. Nursing wages were also positively associated with a higher effect of the cost of capital on total cost, perhaps an artifact of Hospitals in London facing higher nursing and capital costs.

Higher capital costs were associated with increases in the unit cost of Critical Care Episodes, possibly due to the capital intensity of this care setting. A smaller effect in the opposite direction was observed with A&E Attendances, which though statistically significant, was slight in magnitude, a 1% increase in the effective borrowing rate associated with a reduction of £37 in the average cost of an A&E attendance.

Table 3.1: Estimation results - Directly Estimated Quadratic Form

Term	**Estimate**	**Std. Error**
Intercept	302,301,208.050	(4,120,656.240)***
A&E Attendance	234.931	(40.114)***
Critical Care Episode	3,232.185	(183.919)***
Diagnostic Services	2.599	(0.844)**
Elective Inpatient Episode	1,689.908	(140.062)***
Non-Elective Inpatient Episode	679.698	(81.774)***
Outpatient Appointment	90.728	(11.343)***
Training Grade Doctor Wages	1,503.385	(495.716)**
Nurse Wages	5,352.794	(924.745)***
Cost of Capital	936,270.974	(653,431.459)
0.5 * A&E Attendance * A&E Attendance	-0.001	(0.001)
A&E Attendance * Critical Care Episode	-0.012	(0.003)***
A&E Attendance * Diagnostic Services	0.000	(0.000)*
A&E Attendance * Elective Inpatient Episode	0.001	(0.004)
A&E Attendance * Non-Elective Inpatient Episode	0.004	(0.002)
A&E Attendance * Outpatient Appointment	0.000	(0.000)
0.5 * Critical Care Episode * Critical Care Episode	-0.080	(0.009)***
Critical Care Episode * Diagnostic Services	0.000	(0.000)
Critical Care Episode * Elective Inpatient Episode	0.020	(0.012)
Critical Care Episode * Non-Elective Inpatient Episode	0.014	(0.006)*
Critical Care Episode * Outpatient Appointment	0.004	(0.001)***
0.5 * Diagnostic Services * Diagnostic Services	0.000	(0.000)**
Diagnostic Services * Elective Inpatient Episode	0.000	(0.000)
Diagnostic Services * Non-Elective Inpatient Episode	0.000	(0.000)*
Diagnostic Services * Outpatient Appointment	0.000	(0.000)
0.5 * Elective Inpatient Episode * Elective Inpatient Episode	-0.029	(0.018)
Elective Inpatient Episode * Non-Elective Inpatient Episode	0.004	(0.008)
Elective Inpatient Episode * Outpatient Appointment	-0.002	(0.001)
0.5 * Non-Elective Inpatient Episode * Non-Elective Inpatient Episode	-0.010	(0.007)
Non-Elective Inpatient Episode * Outpatient Appointment	0.000	(0.001)
0.5 * Outpatient Appointment * Outpatient Appointment	0.000	(0.000)
A&E Attendance * Training Grade Doctor Wages	-0.010	(0.014)
Critical Care Episode * Training Grade Doctor Wages	0.045	(0.051)
Diagnostic Services * Training Grade Doctor Wages	0.000	(0.000)
Elective Inpatient Episode * Training Grade Doctor Wages	-0.058	(0.055)
Non-Elective Inpatient Episode * Training Grade Doctor Wages	0.054	(0.032)
Outpatient Appointment * Training Grade Doctor Wages	0.003	(0.005)
A&E Attendance * Nurse Wages	0.046	(0.022)*
Critical Care Episode * Nurse Wages	-0.089	(0.046)

R^2 = 0.9817 adj R^2 = 0.9804 AIC = 29,405 BIC = 29,667 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

Table 3.1: Estimation results - Directly Estimated Quadratic Form

Term	**Estimate**	**Std. Error**
Diagnostic Services * Nurse Wages	-0.001	(0.000)
Elective Inpatient Episode * Nurse Wages	0.349	(0.065)***
Non-Elective Inpatient Episode * Nurse Wages	-0.096	(0.041)*
Outpatient Appointment * Nurse Wages	-0.005	(0.006)
A&E Attendance * Cost of Capital	-36.837	(13.818)**
Critical Care Episode * Cost of Capital	238.069	(54.892)***
Diagnostic Services * Cost of Capital	0.647	(0.347)
Elective Inpatient Episode * Cost of Capital	-1.135	(60.384)
Non-Elective Inpatient Episode * Cost of Capital	36.382	(34.546)
Outpatient Appointment * Cost of Capital	-11.181	(5.775)
0.5 * Training Grade Doctor Wages * Training Grade Doctor Wages	-0.340	(0.230)
Nurse Wages * Training Grade Doctor Wages	0.702	(0.323)*
0.5 * Nurse Wages * Nurse Wages	0.572	(0.629)
Cost of Capital * Training Grade Doctor Wages	-372.352	(199.093)
Cost of Capital * Nurse Wages	1,039.379	(332.730)**
0.5 * Cost of Capital * Cost of Capital	219,203.714	(312,778.308)
Teaching Hospital Organisation	-2,145,507.722	(2,757,286.238)

R^2 = 0.9817 adj R^2 = 0.9804 AIC = 29,405 BIC = 29,667 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

Predicted costs fit well to the observed values, with a close association each year between the actual and fitted values, shown in Figure 3.1. In common with the translog specification, the high R^2 value indicates that the estimation may overfit the data and is difficult to apply to data outside the range covered in the sample.

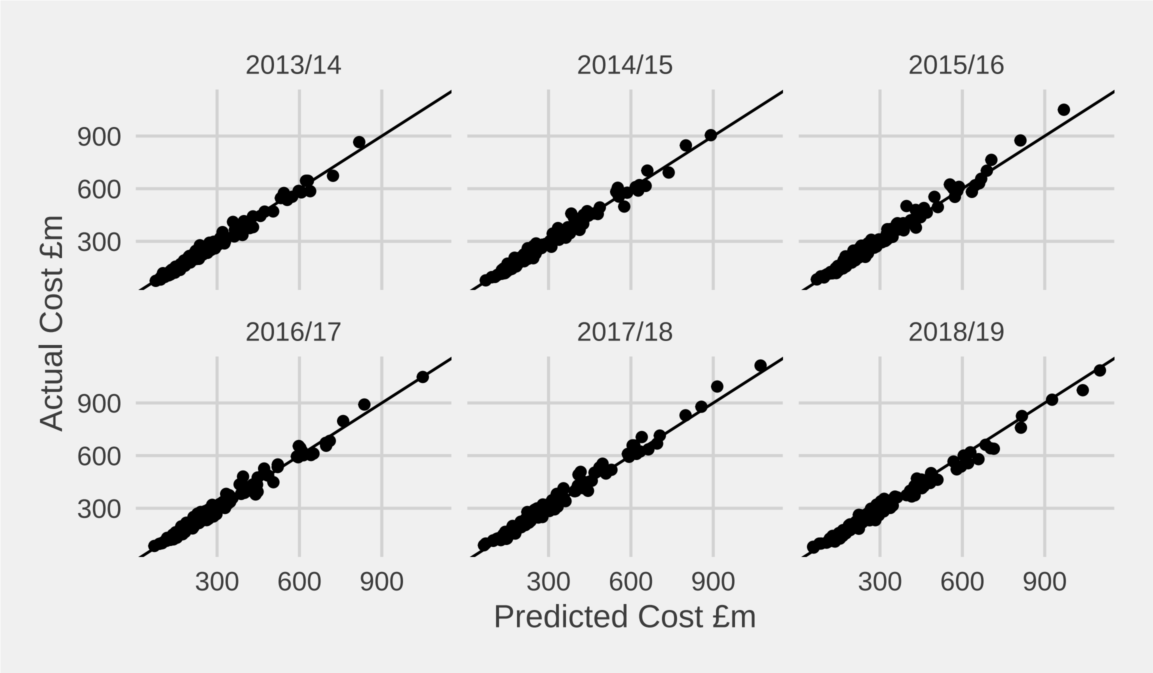

Figure 3.1: Predicted Vs Actual Cost in each Year (Quadratic). Predicted costs are closely correlated to actual values. Predictions are derived from the estimated cost function and actual output levels

3.5.2 Marginal Effects

Taking the partial derivative of total cost with respect to each output, the marginal costs can be calculated in each case. As input and output variables were centred prior to estimation, these marginal costs are interpreted as changes from mean output levels.

As an example, taking the partial derivative with respect to A&E attendances, $\frac{\partial C}{\partial y_{ae}}$ yields:

$$\begin{aligned}
&\mathbf{235} - 0.001 * \mathit{A\&E\,attendances}\, \boldsymbol{-0.012 * Critical\,Care\,Episodes} \\
&+ \boldsymbol{0 * Diagnostic\,Services} + 0.001 * \mathit{Elective\,Inpatient\,Episodes} \\
&+ 0.004 * \text{Non-Elective } \mathit{Inpatient\,Episodes} \\
&+ 0 * \mathit{Outpatient\,Appointments} - 0.01 * \mathit{Training\,Grade\,Doctor\,Wages} \\
&+ \boldsymbol{0.046 * Nurse\,Wages{-}36.837 * Cost\,of\,Capital}
\end{aligned}$$

Terms found to be statistically significant at the 5% level in the estimation are marked in bold.

The constant term, which can be interpreted as the marginal cost of an A&E attendance where all variables are at the mean, is £235. This cost level is comparable with the payment tariff range of £77-359 and the average cost in the dataset as a whole (£156). Though the effect size is small, marginal costs also decrease slightly with increases in the number of critical care episodes carried out. Hospitals with larger critical care units may conduct more costly diagnostics or treatments after admission rather than having diagnostics in A&E. Costs for these activities would then be recorded under the critical care episode rather than the A&E attendance. Additional diagnostic services increase the expected marginal cost of an A&E attendance. Though the effect is statistically significant, the size of the coefficient is small (0.00004), so it is not economically significant. Hospitals that supply high levels of unbundled diagnostic services from non-A&E patients may mean patients attending A&E spend more time in the department waiting for their scans. The marginal cost of an A&E attendance also increases where nurse wages are higher, as higher wages for nursing labour increases the cost of producing a unit of A&E healthcare. Somewhat counterintuitively, the coefficient on the capital price is negative. A&E attendances may be relatively labour-intensive, so they are less affected by increases in capital cost than other outputs.

Taking the partial derivative of cost with respect to critical care episodes gives:

$$\begin{aligned}
&\boldsymbol{3232 - 0.08 * Critical\,Care\,Episodes - 0.012 * A\&E\,attendances}\\
&+ 0 * Diagnostic\,Services + 0.02 * Elective\,Inpatient\,Episodes\\
&\boldsymbol{+ 0.014 *}\ \textbf{Non-Elective}\ \boldsymbol{Inpatient\,Episodes}\\
&\boldsymbol{+ 0.004 * Outpatient\,Appointments}\\
&+ 0.045 * Training\,Grade\,Doctor\,Wages - 0.089 * Nurse\,Wages\\
&\boldsymbol{+ 238.069 * Cost\,of\,Capital}
\end{aligned}$$

The coefficient on the constant term (£3,232) is higher than the mean critical care cost (£1,067), though the critical care term is both statistically significant and negative (-0.08). This result suggests that, excepting interaction effects, scope effects and other factors, additional critical care episodes reduce average critical care costs. As mentioned above, critical care episodes and A&E attendances appear to have economies of scope when colocated. Where hospitals have higher numbers of non-elective care episodes, marginal critical care costs tend to be higher, perhaps because hospitals that provide higher levels of non-elective care are likely to also deal with patients requiring a more costly critical care stay. Similarly, hospitals supplying higher levels of outpatient care may deal with higher requirements for follow-up and rehabilitation after an inpatient stay involving critical care. Higher capital costs increase marginal costs for critical care episodes, possibly due to critical care being capital-intensive and requiring specialist construction and equipment.

As well as taking derivatives of output variables for marginal costs of additional outputs, the partial derivatives of input prices can be used to evaluate the effects of marginal price changes. For example, taking the partial derivative of cost with respect to nurse wages $\frac{\partial C}{\partial w_{nursing}}$ gives:

$$\begin{aligned}
&\boldsymbol{5353 + 0.046 * A\&E\, attendances} - 0.089 * Critical\, Care\, Episodes \\
&- 0.001 * Diagnostic\, Services + \boldsymbol{0.349 * Elective\, Inpatient\, Episodes} \\
&\boldsymbol{-0.096 *}\, \textbf{Non-Elective}\, \boldsymbol{Inpatient\, Episodes} \\
&- 0.005 * Outpatient\, Appointments \\
&+ \boldsymbol{0.702 * Training\, Grade\, Doctor\, Wages} + 0.572 * Nurse\, Wages \\
&+ \boldsymbol{1039.379 * Cost\, of\, Capital}
\end{aligned}$$

Here, the constant term in the derivative (5,353) is broadly in line with the number of staff in the median organisation who would be covered by the same pay conditions as nurses. The relationship between A&E attendances and nursing costs has been addressed above. Hospitals with larger numbers of elective care episodes will have higher costs where the marginal cost of nursing increases. Inpatient care requires higher numbers of nurses than other forms of care, so an increase in nursing wages is likely to disproportionately affect hospitals providing greater amounts of inpatient care, as they require higher numbers of nurses. The sign on non-elective care episodes is the opposite, with higher numbers of non-elective care episodes being associated with a reduction in the marginal effect of nurse wages on total cost. This finding may result from non-elective cases often having shorter hospital stays, so the impact of nursing wages on costs declines as these outputs increase. Hospitals with higher capital costs also have higher marginal effects from increased nursing wages, perhaps because trusts in London tend to have a higher capital cost and pay higher nursing wages.

Similarly, with regard to capital costs, taking the partial derivative $\frac{\partial C}{\partial w_{capital}}$ gives an expression for the marginal cost of capital as follows:

$$\begin{aligned} &936271\boldsymbol{-36.837*A\&E\,attendances+238.069*Critical\,Care\,Episodes}\\ &+0.647*Diagnostic\,Services+-1.135*Elective\,Inpatient\,Episodes\\ &+36.382*\text{Non-Elective }Inpatient\,Episodes\\ &-11.181*Outpatient\,Appointments\\ &-372*Training\,Grade\,Doctor\,Wages+\boldsymbol{1039*Nurse\,Wages}\\ &+219204*Cost\,of\,Capital\end{aligned}$$

Here, the counter-intuitive finding is the opposite signs on the coefficients for interactions with Training Grade Doctor Wages and Nurse Wages. Hospitals that pay higher than average nursing wages experience larger cost increases when the price of capital increases. It is conceivable that hospitals paying higher nursing wages require more nursing labour and capital for inpatient wards. Consequently, an increase in the cost of capital is likely to affect these hospitals disproportionately as they have higher amounts of capital in use. Increases in Training Grade Doctor wages reduce the cost effects of increases in capital costs. Hospitals with higher training grade doctor wages are likely to offer specialty training in addition to foundation programs and could find it easier to substitute medical labour for capital where capital prices are high.

3.5.3 Desirable Properties of the Cost Function

The regression results can be evaluated against the desirable properties of cost functions set out in Chapter 1 :

1. **Cost should be non-negative** — Predicted costs are non-negative for all organisations in the sample. The large intercept value will prevent small activity values from leading to a negative cost, though it is worth noting that the high R^2 value suggests that the model is overfitted and may not represent out-of-sample values well.

2. **Cost should be non-decreasing in input prices** — As wages and prices increase *ceteris parabis*, so should total cost. The partial derivative of cost (or log cost) with respect to each input price (the optimal cost share) ought to be positive. Checking the signs on observed cost-minimising cost shares, 55% of training grade doctor cost shares are negative, 57% of nurses, and 44% of capital prices. These are less congruent with standard theory with the translog results. The translog specification had negative cost shares for

6% observations of junior doctors, 12% nurses, and 53% capital prices. Cost shares (the partial derivative of log cost with respect to log input price) are expected to be positive. The negative cost shares show observations where this desirable property is violated.

3. **Cost should be continuous and concave in input prices** — It is easier to evaluate concavity with input prices with the quadratic specification, as the diagonal elements of the input prices are equal to the coefficient on the square term multiplied by the input price. The input price being positive, negativity is therefore determined by the sign on the coefficient. In this case, only training grade doctors have a negative coefficient on the squared term, with the coefficients for squared nurse pay and the capital price both being positive. Nursing wages and capital prices are, therefore, not concave with respect to cost, so it cannot be said that the cost function as a whole is concave with respect to input prices. For the translog specification, the calculation was more complex, with 798 / 798 observations being concave with respect to training grade doctor wages, 0 / 798 observations concave with respect to nursing wages, and 48 / 798 capital price observations. The translog specification is consequently closer to being concave with respect to input prices.

4. **Cost should be non-decreasing in any output** — This property can be evaluated by looking at the individual cost-output elasticities calculated for each output category. Where these elasticities are negative, a percentage increase change in output causes a percentage decrease in cost. There are 6 output categories and 798 observations, giving 4788 total cost-output elasticities calculated. Negative cost-output elasticities are present in 66 / 798 A&E observations, 4 / 798 critical care observations, 95 / 798 diagnostic observations, 25 / 798 elective inpatient observations, 6 / 798 non-elective inpatient observations, and 13 / 798 outpatient appointment observations. Apart from A&E and diagnostic services, there are few cases of non-compliance with this desired characteristic. Cost-output elasticities for A&E and diagnostic services are more likely to be negative than other outputs at particular observations as these elasticities are generally lower than other outputs. Their low level may be either because their outputs form a lower share of total costs or because they have low marginal costs. Consequently, local violations of cost function convexity are more likely to occur in these outputs, and these results are less concerning. These results can be compared to the translog specification, which had 49 / 798 negative elasticities in A&E outputs, 7 / 798 in A&E critical care outputs, 55 / 798 in diagnostic service outputs, 0 / 798 in elective inpatient episodes, 0 / 798 in non-elective inpatient episodes, and 0 / 798 amongst

outpatient outputs.

5. **Cost should be linearly homogeneous in input prices** - In the translog form, taking logs of input prices ensures that the resulting coefficient can be interpreted as the percentage change in total cost for a 1% increase in the coefficient value. If the sum of coefficients for (non-interacted) input prices is equal to one, and the sum of cross-product terms for each output is zero, then percentage increases in all input prices yield a proportionate percentage increase in total cost. This property was not true of the translog cost function estimated in Chapter 2. It is theoretically possible to enforce this via a constrained regression for the translog, but in the quadratic form, values are not logged, so no equivalent interpretation exists. In contrast to the translog specification, It is not possible in the quadratic case to *a priori* ensure the estimated cost function is homogeneous of degree 1. A 1% rise in total input prices in the quadratic form would raise the total cost by different percentages depending on the coefficients estimated on input price terms.

6. **No fixed costs in the long-run** - The large intercept term observed in the estimation results precludes the estimation returning zero costs where all inputs are zero. This phenomenon was also a feature of the translog estimation.

Comparing the results above with the translog format, both functional forms meet criterion 1. The translog specification is closer to the desired properties than the estimated quadratic for criteria 2, 3 and 4 and can also be made compliant with criterion 5. Neither estimated forms meet criterion 6 due to the inclusion of an intercept term in the estimation. According to the above, there are grounds to prefer the translog as more consistent with economic theory.

Both specifications are more consistent with respect to outputs than input prices. Input prices have more departures from concavity and areas where price increases are not associated with cost increases. This observation is particularly the case for the capital price, suggesting that the construction of the weighted cost of capital may struggle to fully capture variations in the cost of capital. The relatively worse performance of input variables compared to output variables is also likely a function of missing factors for which adequate data is unavailable, for example, agency staff and consumables. Consequently, coefficients on input prices and their cross-products may have high standard errors and are more vulnerable to omitted variable bias than output variables.

3.5.4 Marginal and Average Cost Visualisations

I visualise computed marginal and average costs below, using the same method as Chapter 2. In contrast to the translog cost curves, the quadratic marginal and average cost curves are linear rather than curved. The linear shape follows the quadratic nature of the specification. Taking a partial derivative of a quadratic function to define a marginal cost leaves an expression where no term has a higher power than one. Cost curves shown below are constructed by holding all factors other than one output constant. Consequently, marginal and average cost curves are linearly related to that particular output. The shape of marginal and average cost curves for the quadratic and translog specification is worthy of comment. The log terms in the translog specification gave convex average and marginal cost curves when computed for each observation. For the quadratic specification, taking the partial derivative gives a linear expression for both average and marginal costs, holding all but one output constant. As scale economies are defined as the ratio of average costs to marginal costs, a quadratic specification where changes to scale economies are linear may be less sensitive to changes in small outputs than a translog specification. For a translog specification, small absolute output changes amongst low-output hospitals may represent large percentage increases in output (and cost), meaning the ratio of average cost and marginal costs could be more volatile.

It is possible to compare these to the translog specification to look at differences in how costs vary with output. Figure 3.2 shows the marginal costs of elective inpatients for a selection of 30 observations. The slope of the marginal cost curves is steeper than the translog specification. Costs decline more with increases in output, in contrast with the translog form that showed marginal costs of elective inpatient work being flatter. The relationship also holds with average costs in Figure 3.5. Scale elasticities in elective work are accordingly higher in the quadratic specification than the quadratic one. Non-elective marginal and average costs are flatter using the quadratic specification than were observed in the translog. Costs declined more sharply with increases in output for the translog, whereas the curves are less steep for the quadratic. Outpatient marginal and average costs are relatively flat in both translog and quadratic specifications.

Comparing the two sets of visualisations, generated marginal and average cost curves were of a comparable level for both the translog and quadratic forms. Observed average costs were in line with the mean values for the dataset as a whole, set out in Table 2.3.

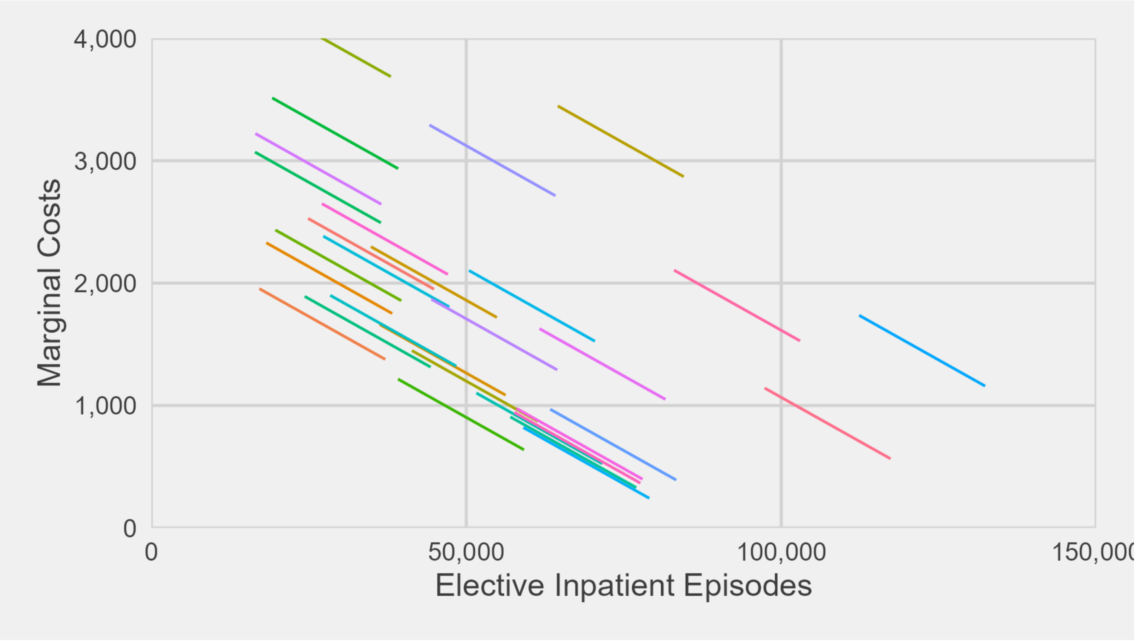

Figure 3.2: Elective Inpatient Episode Marginal Cost Curves (Quadratic). Marginal cost curves are derived from the estimated cost function by calculating the marginal cost effects of incremental changes in one output type, holding other outputs constant. The plot shows sections of cost curves generated from a random sample of 30 observations, differentiated by colour

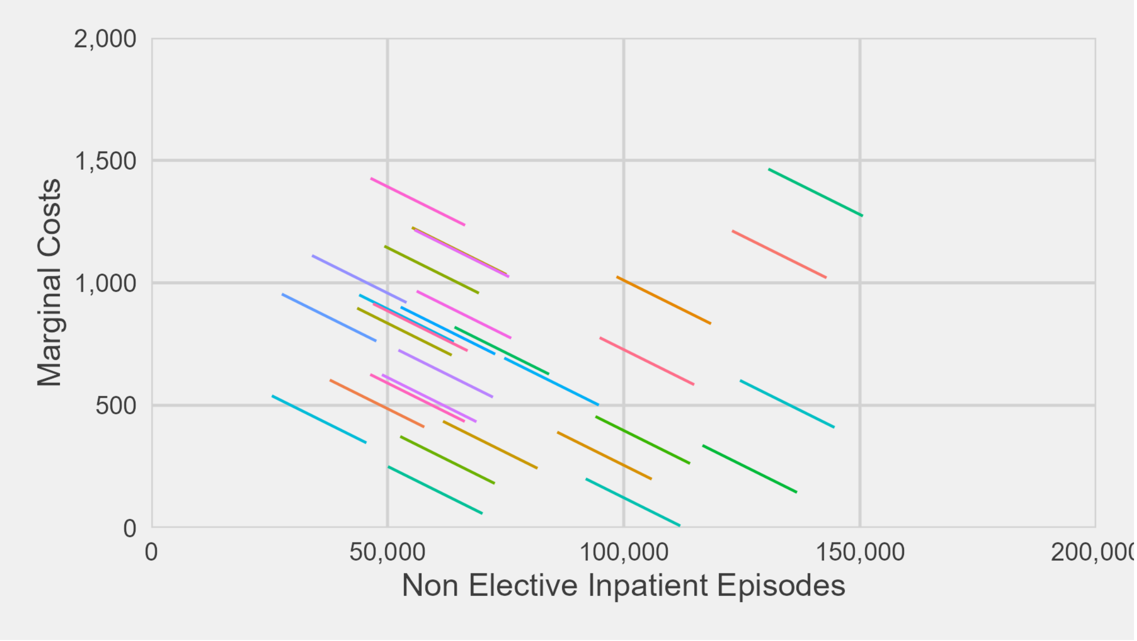

Figure 3.3: Non-Elective Inpatient Marginal Cost Curves (Quadratic). See Figure 3.2 for details on derivation.

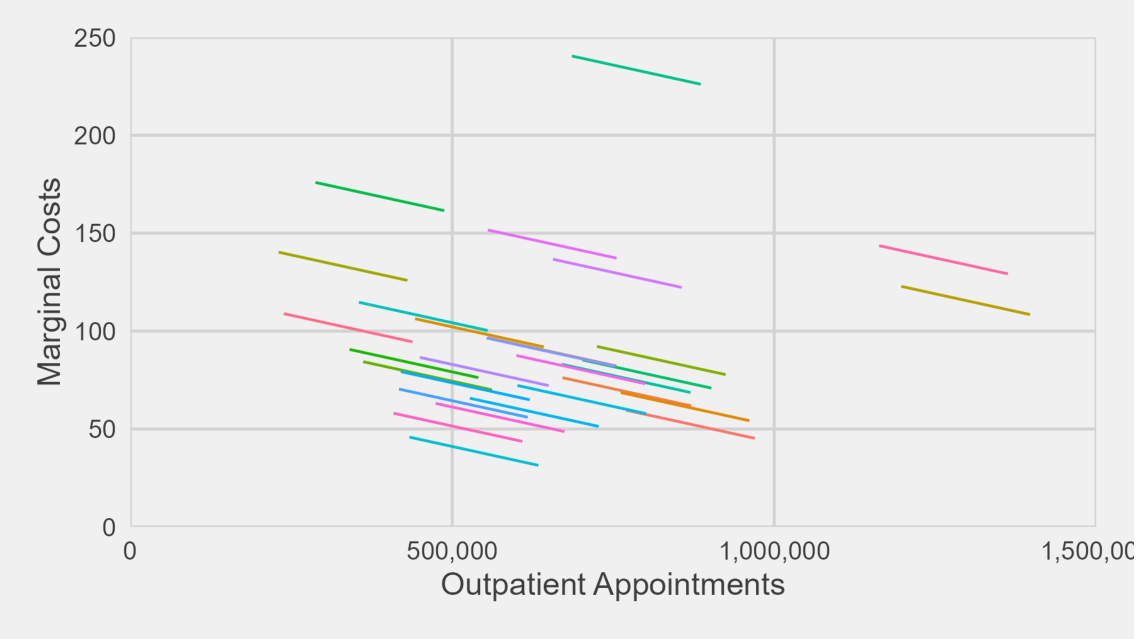

Figure 3.4: Outpatient Marginal Cost Curves (Quadratic). See Figure 3.2 for details on derivation. Outpatient appointment costs are relatively invariant to output levels. Appointments are a discrete period of time in consultation with a doctor with consequent limitations on scale economies.

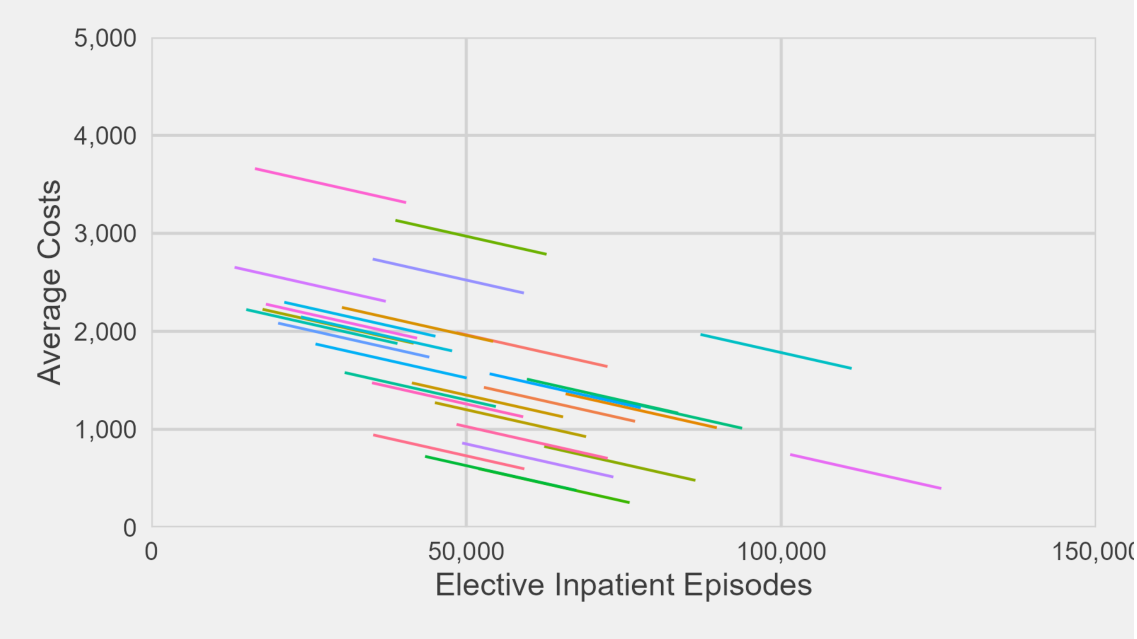

Figure 3.5: Elective Inpatient Average Cost Curves (Quadratic). Average Cost Curves are derived from the estimated cost function by calculating the total cost effects of discrete changes in one type of output and then dividing by the change in output, holding other outputs constant. The plot shows sections of cost curves generated from a random sample of 30 observations, differentiated by colour

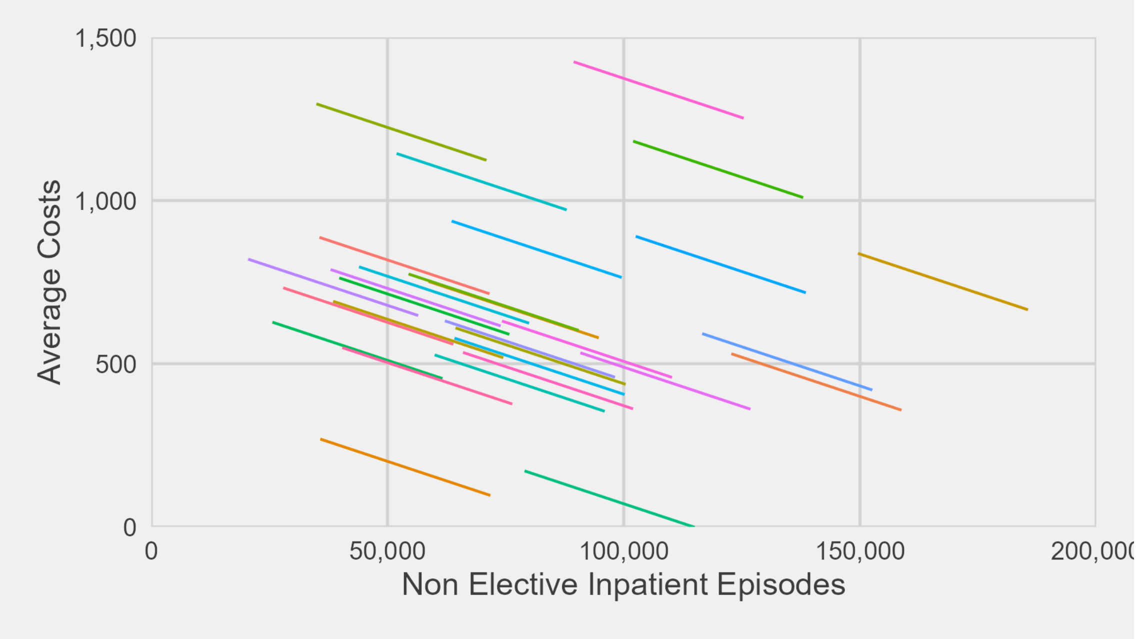

Figure 3.6: Non-Elective Inpatient Average Cost Curves (Quadratic). See Figure 3.5 for details on derivation.

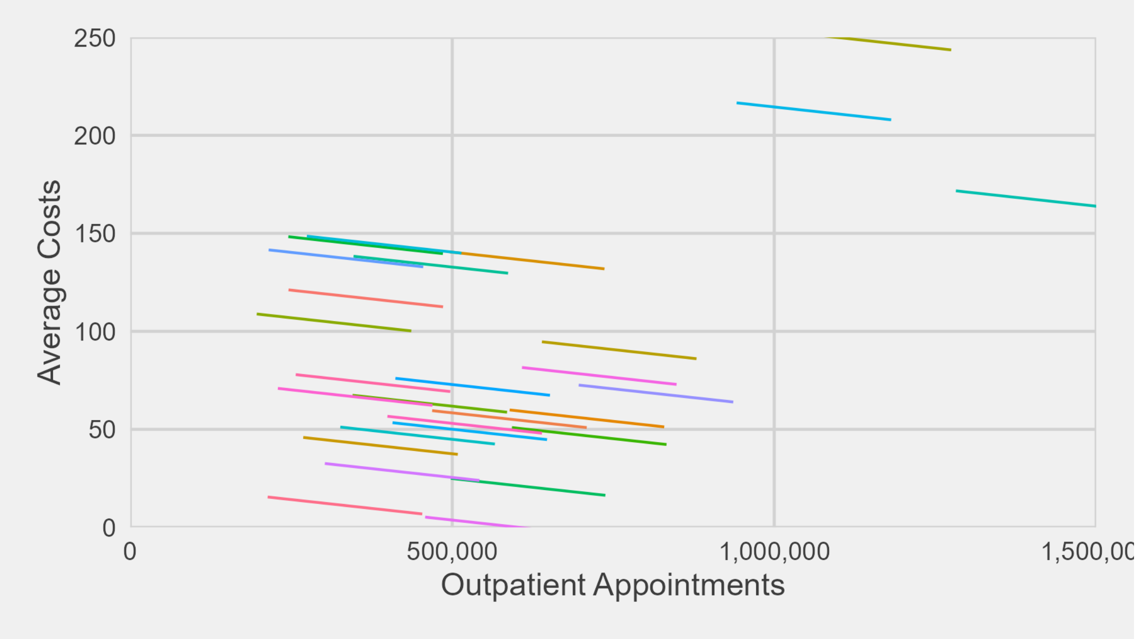

Figure 3.7: Outpatient Average Cost Curves (Quadratic). See Figure 3.5 for details on derivation. Outpatient appointment costs are relatively invariant to output levels. Appointments are a discrete period of time in consultation with a doctor with consequent limitations on scale economies.

3.5.5 Scale Economies

The scale economy estimate for each organisation remains relatively consistent over the years, as was the case for the translog estimation, with few large changes year on year for each organisation.

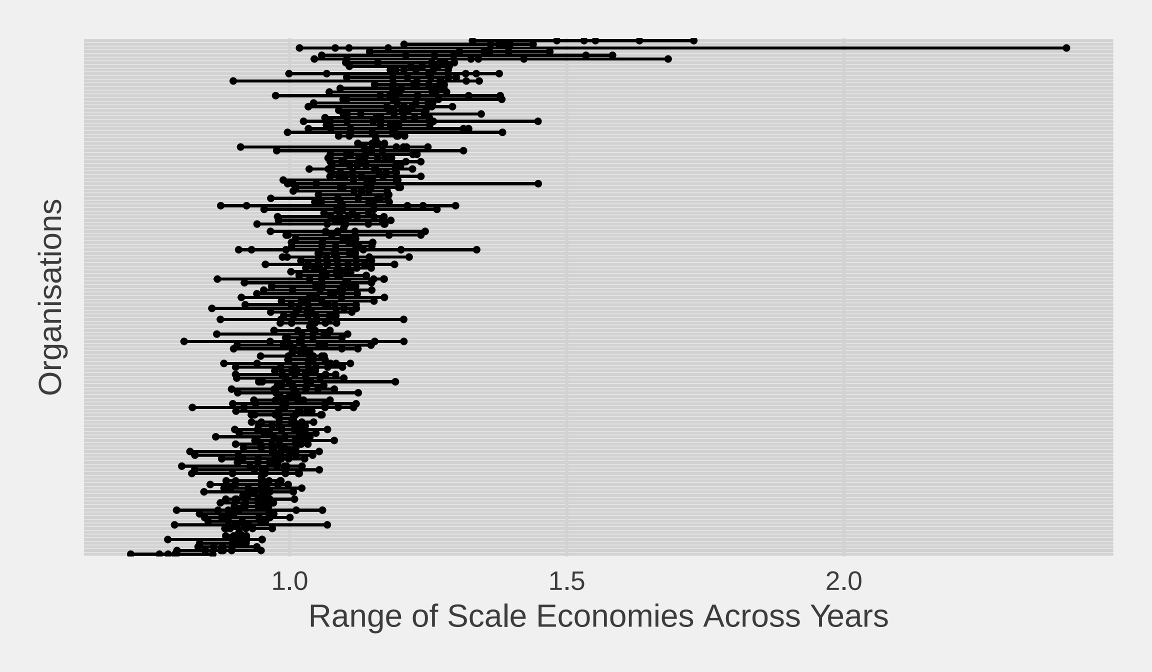

Figure 3.8: Range of Scale Economy Indices by Organisation (Quadratic). As with the translog form, scale economies are relatively stable for each organisation across the period

Scale economy estimates are less related to size compared to the translog specification. Figure 3.9 plots scale economy measures against size as measured by the number of beds in 2017-18. Most hospitals have constant economies of scale, but scale economies are not related to the number of beds. A polynomial fitted to the data has a broadly flat shape showing no apparent limit to scale economies over the dataset range. Expanding or reducing scale does not affect the scale economy calculation, so hospitals, irrespective of size, would see similar effects of expansion at all sizes.

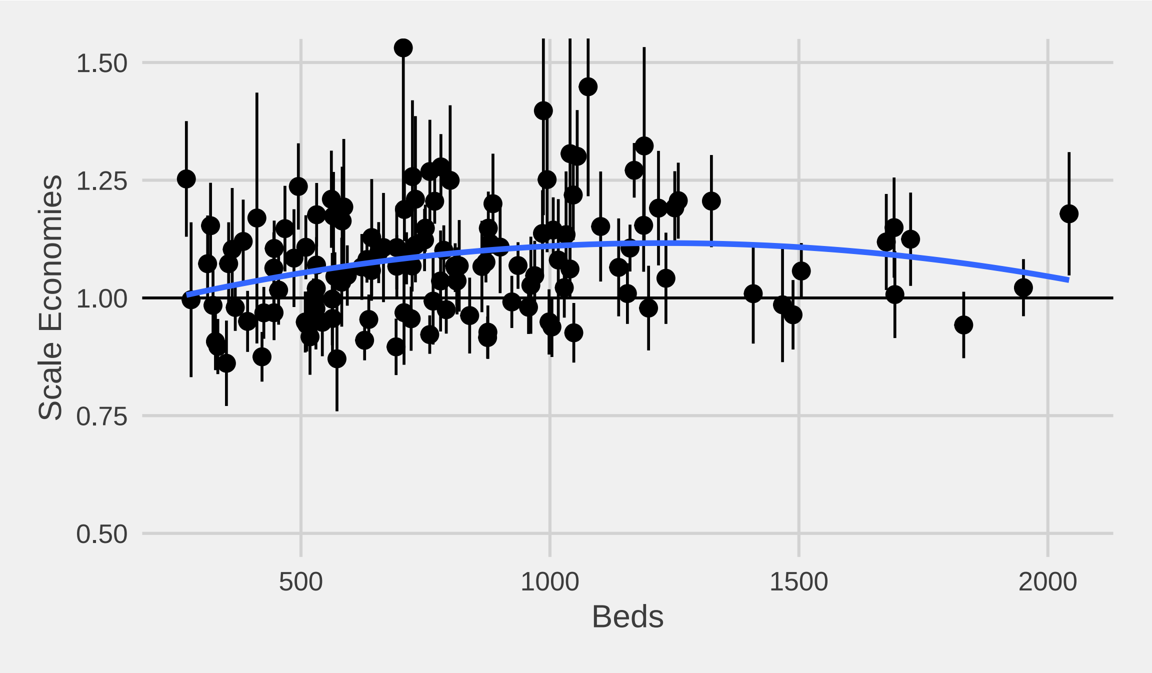

Figure 3.9: Scale Economy Index Vs Bed Numbers (Quadratic) - 2017-18 Observations only. Here the year 2017-18 is shown rather than the most recent year in the dataset, as scale economies for that year appear to be affected by data suppression

The relationship between scale economies and size was consistent over the years, though estimated scale economies were slightly lower in 2018/19 than in other years. This year has slightly less activity than other years as new data suppression rules censored more data from the dataset. These censor the long tail of very infrequently recorded HRG codes where case numbers were below 8 for the year. Low numbers are suppressed to minimise risks of identification with released data. Suppression was relatively higher in elective and non-elective care, where episode numbers are lower overall and therefore have more categories where activity numbers were below the threshold. Admitted care, particularly non-elective care, appears to have higher scale economies than other care settings, so the change in the censoring may lower calculated scale economies overall. This feature does not appear to be present in the translog specification computed in Chapter 2, perhaps because the standard errors on the point estimates are lower for the translog, and there is less volatility overall.

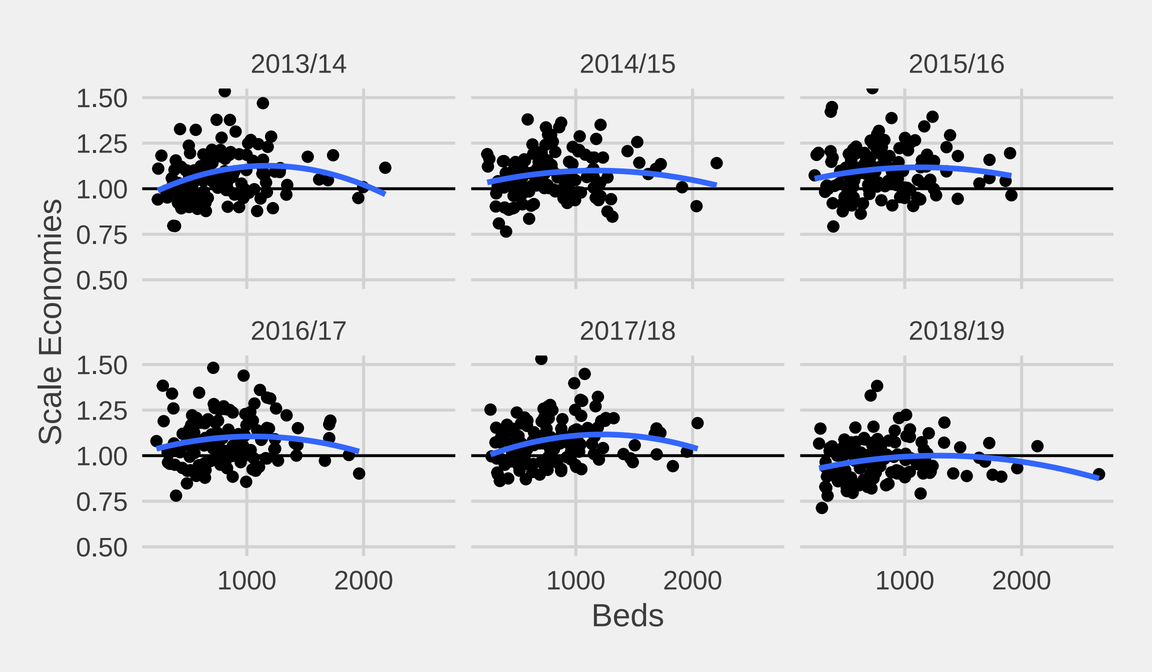

Figure 3.10: Scale economies vs Bed numbers across Time (Quadratic). Values are consistent apart from 2018/19, which is lower due to additional small number suppression in that year

Though the range of scale economies is compressed relative to the translog estimation, regional differences observed a similar pattern, with more rural or lower-wage regions showing slightly more scale economies than higher-wage London.

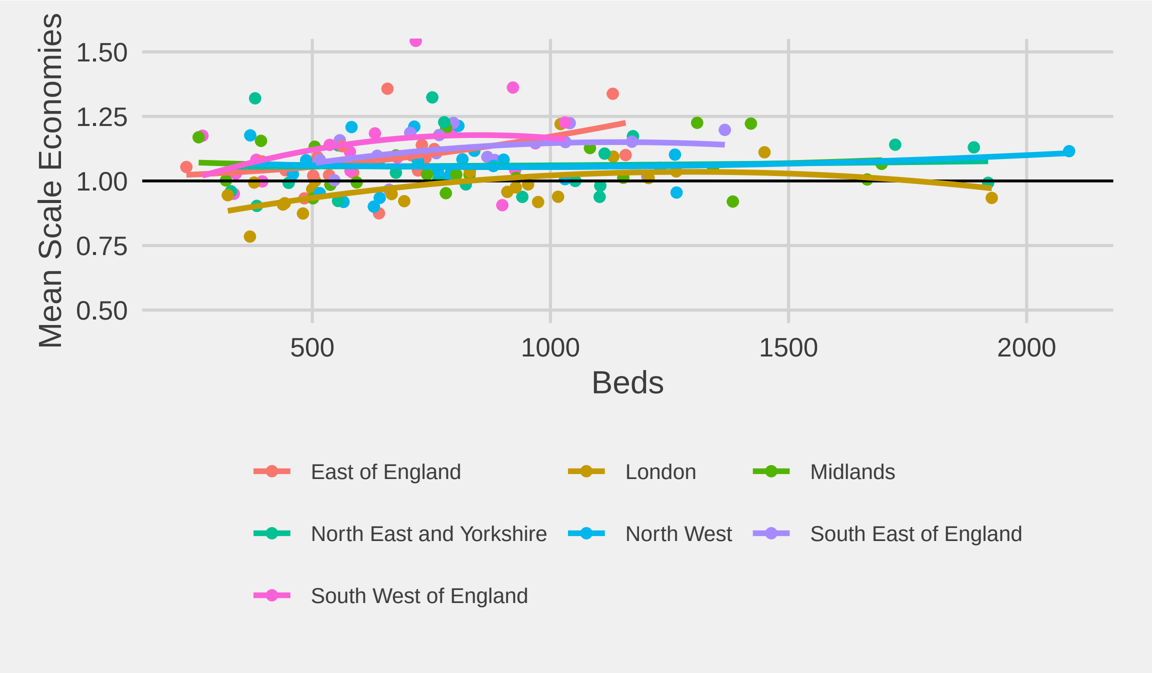

Figure 3.11: Scale Economies, by NHS England region (Quadratic). Organisations in London and neighbouring areas in the South East are required to pay higher wages in the form of a wage supplement for high-cost areas. This supplement is intended to compensate staff required to work in higher cost areas, effectively raising variable costs for organisations in London or proximate to it. The quadratic form displays a similar pattern to the translog form, with London having lower scale economies than elsewhere

3.6 Discussion

Direct estimation of the cost function shows broadly comparable results for the translog and quadratic forms. The main coefficients on output and input prices are comparable with average costs in the case of output variables, and the average number of staff in the case of the input variables. The large number of interaction terms in the specification implies caution in interpretation of individual coefficients, but there are suggestions that critical care unit costs may decline with output, but are associated with higher costs in other settings, such as Non-Elective Inpatient and Outpatient Appointments. Unit costs of activity were more affected by nursing than medical wages, particularly in Elective Inpatient care and in A&E.

When used to derive scale economy estimates, the translog showed small but positive scale economies up to around 1,200 beds. The quadratic estimation also shows small and generally positive scale economies, with the median scale economy index being 1.05. These results are comparable with estimates shown in Chapter 2. However, the relationship to size as defined by bed numbers is less clear, with scale economies showing no obvious relation to size, and no apparent limit to scale economies was observed. Under a quadratic specification, larger trusts still record similar scale economies to smaller trusts, suggesting that the limit obtained in Chapter 2 is sensitive to the form of the cost function used.

If a choice between the two functional forms is necessary, there are some theoretical reasons that the translog is preferable in this case. Input prices were more consistent with expected theoretical properties, with fewer departures from concavity. The translog estimates also had narrower standard errors on scale economy calculations for each observation, giving more precise individual estimates. However, there are few general reasons to prefer one form over the other.

This chapter extends the analysis from Chapter 2 to consider an alternate functional form. Quadratic functional forms are typically less prevalent in the literature, and few studies contrast multiple functional forms. Computing results from the same data using different forms allows us to evaluate the effects of particular functional forms. Studies that report results using only one form may omit uncertainty arising from using only one specification. Estimating limits of minimum efficient scale may be particularly affected by the choice of form if the ends of the size distribution are affected in different ways.

Policymakers may take some assurance from these results that scale economies in English hospitals are small but generally positive and that whilst results appear variable based on functional form, most hospitals in the middle of the distribution have comparable results. There are some areas where gains could be realised, though changes to hospital healthcare supply would be subject to equity and access considerations. Policymakers may need to be aware that computed limits to scale economies may be hard to quantify where there are few observations.

Future studies could consider using both forms to quantify uncertainty arising from different functional forms, or at least compare how well each form fits the data, justifying the particular functional form selected rather than using one form only. Studies that do attempt to quantify limits to scale economies should compute results using multiple functional forms. Using multiple forms may be particularly useful where scale economies are low or do not vary much with size. Future studies could also investigate why differences are observed between translog and quadratic functional forms, whether the patterns identified here are generally true, and whether there are any reasons to suppose one functional form is preferable to the other for scale economy studies.

Chapter 4

Economies of Scale in English Hospitals: Estimation via a Short-run Cost Function

4.1 Introduction

The previous analyses in Chapter 2 and Chapter 3 directly estimate long-run cost functions from organisations and observed input prices. Directly estimating a cost function from observed values implicitly assumes that these values are optimal. This assumption may not be justifiable in the scenario where the capital stock has not been fully adjusted to current output, or there are other factors inhibiting cost-minimisation.

This chapter addresses this issue by using an envelope methodology, estimating a short-run translog cost function including capital levels proxied by bed numbers. The short-run function is then solved for optimal capital levels as described above before substituting back into the cost function to yield a long-run cost function. This way, the assumption of cost-minimisation and optimal capital levels at the observed cost and output levels can be relaxed. The same envelope approach is used with a quadratic cost function to see how this approach affects that functional form.

Results show that scale economies estimated using an envelope methodology are higher in both the short and long run, as would be expected if currently observed capital levels were not optimal. However, optimal capital values were observed to be sensitive to the cost function's functional form and specification. Using beds as a proxy for capital may not adequately capture underlying capital levels and consequently may not accurately adjust for hospitals not on the frontier of the long-run cost function. Future studies should use this approach carefully and with adequate sensitivity analysis.

4.2 Previous Literature

Several authors, starting from Cowing & Holtmann (1983) and Vita (1990), note that direct estimation assumes each observation is the lowest possible cost for that particular combination of outputs. Observations would only be at that point if capital allocations were optimal at the point of measurement. These studies argue that there is no guarantee that currently observed capital levels are optimal. Departures from non-optimal capital levels could arise due to technological change, for example, if hospitals have not been able to invest in new capital innovations or have capital endowments that have become obsolete. Alternatively, hospitals may have been unable to adjust capital in time to match their current level of healthcare demand. Mismatched capital levels may increase expenditure. For example, a hospital may need additional buildings and facilities to deal with an increase in demand in a cost-optimal way, but it may not be able to construct those buildings in the short-run. To deal with the issue in the short-run the hospital may need to adopt a more expensive labour-intensive production technology. In the long-run, the hospital may be able to construct additional buildings and produce at a lower cost, moving down to an optimal position on the cost curve. Using current output and capital levels in the estimation may misestimate potential scale economies by including non-optimal average cost points in the final dataset.

Alternatively, hospitals may not be fully cost-minimising in the short-run, especially where management may have clinical or growth objectives that do not fit into the paradigm of the cost-minimising neoclassical firm. Hospitals may also lack incentives to minimise costs where they have no local competition. These are specific instances of the more general phenomenon of X-inefficiency (Leibenstein, 1966) that could be present in observed cost values. Such considerations mean that the observed cost profile may not be fully cost-minimising, and again, directly estimating long-run cost functions from observed production values may understate scale economies.

To deal with these issues, some authors, including Cowing & Holtmann (1983), restrict the estimation to short-run cost functions only. Other authors develop their study by generating long-run functions from the short-run estimations rather than estimating the long-run directly (Aletras, 1999; Kristensen et al., 2012; Preyra & Pink, 2006; Vita, 1990). These authors follow a common methodology of first estimating a short-run cost function. This short-run function will include fixed factors of production (usually capital as proxied by beds) when the estimation is carried out. An envelope of the short-run function is then taken by first differentiating the short-run function with respect to capital, setting the result equal to zero, and then solving for optimal capital levels in terms of outputs. The resulting optimal capital relationship can be substituted back into the short-run function to yield a long-run cost function, allowing for cases where actual capital levels are non-optimal. Following Cowing and Holtman, who used a translog form, Preyra & Pink and Kristensen et al. use a quadratic form to estimate a short-run function before computing the envelope to derive the long-run cost function. These studies typically find that taking the envelope of short-run cost functions returns a higher estimate of scale economies than direct estimation of the cost function from observed values.

4.3 Data

This chapter uses the same cost and activity data used in previous chapters and combines this with data on bed numbers and financial statement asset values. The additional data is used to construct proxy measures for capital at each organisation. The relative advantages and disadvantages of each as a proxy for capital are discussed in detail in Section 1.4.7. Both bed numbers and financial statement values have significant drawbacks as proxy measures of capital. However, in the analysis, beds are used as the preferred measure of capital. This decision is due to the inherent variability of financial statement capital accounting and the artificial nature of hospital property valuation in the absence of a developed market for hospital buildings.

4.4 Method

This chapter computes a long-run cost function via an envelope of short-run cost functions. Firstly a short revised short-run cost function is estimated, including bed numbers as a proxy variable for the level of the capital stock. The estimation uses a random effects estimator and robust standard errors. The beds variable is interacted with the output and price variables. An envelope of short-run functions is then constructed by solving for optimal capital stock (beds), substituting the expression for optimal capital stock into the short-run cost function. The new expression is a long-run cost function which assumes that capital can be set to the cost-minimising optimal value for that output level.

The sensitivity of results is explored by using slightly different specifications in the translog form, as well as a quadratic functional form, to compare how the results of the envelope method may differ with changes in the cost function specification and form. For each functional form and specification, optimal beds are calculated along with scale economy estimates.

4.4.1 Estimating a Short-run Cost Function

A short-run translog cost function is estimated, including a variable for capital in the specification. Including this capital variable makes the cost function conditional on the level of capital employed. Capital is assumed to be fixed in the short-run. As discussed in Section 1.4.7, bed numbers are used as the preferred proxy for capital employed.

The short-run cost function, using a translog specification, becomes:

$$\begin{aligned}\ln\, C_{ht}(y,w,K') =& \alpha_0 + \sum_i^m \alpha_i \ln\, y_{iht} + \sum_k^n \beta_k\, \ln\, w_{kht} + \frac{1}{2}\sum_i^m\sum_j^m \delta_{ij}\, \ln\, y_{iht}\, \ln\, y_{jht} \\ & + \frac{1}{2}\sum_k^n\sum_l^n \gamma_{kl}\, \ln\, w_{kht}\, \ln\, w_{lht} + \sum_i^m\sum_k^n \rho_{ik}\, \ln\, y_{iht}\, \ln\, w_{kht} \\ & + \frac{1}{2}\zeta\, \ln\,(K'_{ht})^2 + \eta\, \ln\, K'_{ht} + \sum_i^m \theta_i\, \ln\, y_{iht}\, \ln\, K'_{ht} \\ & + \sum_k^n \mu_k\, \ln\, w_{kht} \ln\, K'_{ht} + \epsilon_{ht}\end{aligned} \tag{4.1}$$

Where K' is the level of capital employed, and y, w are the m outputs and n input prices. As before, Hospital index h and time period t reflect the panel data nature of the dataset. The error term $\epsilon_{ht} = \tau_h + v_{ht}$ comprises a term τ_h representing hospital specific effects and a random noise term v_{it}. For reasons of clarity, time and individual subscripts are omitted from this point on in the chapter.

4.4.2 Solving for the Long Run

Short-run costs are a function of variable costs plus the total cost of capital at the current (fixed) capital level. Once estimated, the partial derivative of the short-run cost function with respect to capital is obtained and set to zero:

$$\frac{\partial \ln\, C(y,w,K')}{\partial \ln\, K'} = 0 \tag{4.2}$$

Solving this gives an expression for optimal capital stock K^* in terms of healthcare outputs and input prices. This expression can be substituted back into the short-run results to give a long-run cost function determined by outputs and input prices. This version of the long-run function allows for capital levels to be non-optimal. Essentially this approach creates an envelope of the short-run cost functions, allowing for the case where hospitals are not currently on the frontier of lowest cost given long-run output levels. Computing scale economies from the long-run specification is equivalent to calculating scale economies using the short-run cost function, assuming each hospital has its computed optimal bed number given the current output level. In effect, the long-run cost equation assumes that each hospital can obtain the optimal capital stock at each output level. The true long-run cost function may be lower than that estimated directly from observed values. This weakness of direct estimation could lead to understating limits to scale economies or obtaining point estimates that are distorted by the presence of X-inefficiency or non-optimal capital stock.

4.5 Results

4.5.1 Short-run Estimation

Details of this regression are set out in Table 4.1. With the inclusion of beds in the short-run function, the coefficients on non-interacted outputs and input prices change but retain their statistical significance where they were significant in the direct estimation (Table 2.7). The additional bed variables include cross-products with each type of output and input price, and a squared bed term. The squared bed term and the non-interacted bed term are statistically significant, along with interactions with critical care, non-elective inpatients, and nursing wages. Critical Care activity has a high cost 'per bed' because of the high acuity of care. Non-elective care costs may also be affected by bed numbers as hospitals need a sufficient number of beds to manage peaks in demand for non-elective care. Hospitals with higher levels of non-elective care may require higher bed numbers to manage the higher volatility of demand, which increases average costs where beds are not used. The other statistically significant interaction is between beds and nursing wages. Higher nursing wages are associated with a stronger effect of additional beds on costs. Additional beds require a certain level of additional nursing, so where this nursing is expensive, there will be a higher effect on the total cost.

Table 4.1: Estimation results - Short-run Translog Form

Term	**Estimate**	**Std. Error**
Intercept	19.365	(0.016)***
A&E Attendance	0.064	(0.015)***
Critical Care Episode	0.077	(0.013)***
Diagnostic Services	0.049	(0.008)***
Elective Inpatient Episode	0.273	(0.025)***
Non-Elective Inpatient Episode	0.250	(0.018)***
Outpatient Appointment	0.122	(0.022)***
Training Grade Doctor Wages	0.452	(0.072)***
Nurse Wages	0.462	(0.125)***
Cost of Capital	0.000	(0.008)
0.5 * A&E Attendance * A&E Attendance	-0.081	(0.048)
A&E Attendance * Critical Care Episode	-0.033	(0.032)
A&E Attendance * Diagnostic Services	0.029	(0.031)
A&E Attendance * Elective Inpatient Episode	-0.088	(0.076)
A&E Attendance * Non-Elective Inpatient Episode	0.026	(0.053)
A&E Attendance * Outpatient Appointment	-0.049	(0.066)
0.5 * Critical Care Episode * Critical Care Episode	-0.063	(0.026)*
Critical Care Episode * Diagnostic Services	-0.028	(0.017)
Critical Care Episode * Elective Inpatient Episode	-0.032	(0.057)
Critical Care Episode * Non-Elective Inpatient Episode	0.013	(0.037)
Critical Care Episode * Outpatient Appointment	0.022	(0.046)
0.5 * Diagnostic Services * Diagnostic Services	-0.041	(0.018)*
Diagnostic Services * Elective Inpatient Episode	-0.054	(0.047)
Diagnostic Services * Non-Elective Inpatient Episode	0.050	(0.035)
Diagnostic Services * Outpatient Appointment	0.033	(0.043)
0.5 * Elective Inpatient Episode * Elective Inpatient Episode	0.222	(0.157)
Elective Inpatient Episode * Non-Elective Inpatient Episode	0.016	(0.086)
Elective Inpatient Episode * Outpatient Appointment	-0.151	(0.111)
0.5 * Non-Elective Inpatient Episode * Non-Elective Inpatient Episode	0.001	(0.094)
Non-Elective Inpatient Episode * Outpatient Appointment	0.152	(0.082)
0.5 * Outpatient Appointment * Outpatient Appointment	0.143	(0.125)
0.5 * Average Beds * Average Beds	-0.336	(0.165)*
Average Beds	0.140	(0.027)***
Average Beds * A&E Attendance	0.133	(0.078)
Average Beds * Critical Care Episode	0.202	(0.056)***
Average Beds * Diagnostic Services	0.066	(0.050)
Average Beds * Elective Inpatient Episode	0.141	(0.143)
Average Beds * Non-Elective Inpatient Episode	-0.279	(0.092)**
Average Beds * Outpatient Appointment	-0.160	(0.104)
Average Beds * Training Grade Doctor Wages	0.370	(0.356)

R^2 = 0.9852 adj R^2 = 0.9838 AIC = -2,819 BIC = -2,505 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

Table 4.1: Estimation results - Short-run Translog Form

Term	**Estimate**	**Std. Error**
Average Beds * Nurse Wages	-1.188	(0.494)*
Average Beds * Cost of Capital	-0.048	(0.043)
A&E Attendance * Training Grade Doctor Wages	0.372	(0.244)
Critical Care Episode * Training Grade Doctor Wages	0.281	(0.158)
Diagnostic Services * Training Grade Doctor Wages	-0.087	(0.140)
Elective Inpatient Episode * Training Grade Doctor Wages	-0.280	(0.334)
Non-Elective Inpatient Episode * Training Grade Doctor Wages	-0.169	(0.276)
Outpatient Appointment * Training Grade Doctor Wages	-0.148	(0.321)
A&E Attendance * Nurse Wages	-0.011	(0.301)
Critical Care Episode * Nurse Wages	0.241	(0.195)
Diagnostic Services * Nurse Wages	-0.136	(0.174)
Elective Inpatient Episode * Nurse Wages	0.900	(0.517)
Non-Elective Inpatient Episode * Nurse Wages	-0.673	(0.358)
Outpatient Appointment * Nurse Wages	0.540	(0.477)
A&E Attendance * Cost of Capital	-0.002	(0.028)
Critical Care Episode * Cost of Capital	0.055	(0.016)***
Diagnostic Services * Cost of Capital	0.030	(0.018)
Elective Inpatient Episode * Cost of Capital	-0.042	(0.034)
Non-Elective Inpatient Episode * Cost of Capital	-0.011	(0.027)
Outpatient Appointment * Cost of Capital	-0.001	(0.036)
0.5 * Training Grade Doctor Wages * Training Grade Doctor Wages	-2.932	(1.467)*
Nurse Wages * Training Grade Doctor Wages	-3.883	(1.824)*
0.5 * Nurse Wages * Nurse Wages	5.668	(3.431)
Cost of Capital * Training Grade Doctor Wages	0.023	(0.119)
Cost of Capital * Nurse Wages	0.045	(0.166)
0.5 * Cost of Capital * Cost of Capital	0.048	(0.018)**
Teaching Hospital Organisation	0.024	(0.025)

R^2 = 0.9852 adj R^2 = 0.9838 AIC = -2,819 BIC = -2,505 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

The optimal beds calculation as set out in (4.2) yields:

$$\begin{aligned}\frac{\partial \ln C(y,w,K')}{\partial \ln K'} =& -0.336 \cdot \ln\ Beds + 0.14 + 0.133 \cdot \ln\ A\&E \\ & + 0.202 \cdot \ln\ Critical\,Care\,Episodes + 0.066 \cdot \ln\ Diagnostic\,Services \\ & + 0.141 \cdot \ln\ Elective\,Inpatient\,Episodes \\ & - 0.279 \cdot \ln\ \text{Non-Elective}\ Inpatient\,Episodes \\ & - 0.160 \cdot \ln\ Outpatient\,Appointments \\ & + 0.370 \cdot \ln\ Training\,Grade\,Doctor\,Wages \\ & - 1.188 \cdot \ln\ Nursing\,Wages - 0.048 \cdot \ln\ Cost\,of\,Capital \end{aligned} \tag{4.3}$$

This expression can be considered as the cost elasticity of capital, evaluated at mean values. A 1% increase in capital (beds) is associated with a 0.2% unit cost increase in Critical Care, but a 0.28% decrease in the unit cost of Non-Elective Episodes, as well as a 1.19% decrease in the impact of nursing wage increases on the total cost. Other coefficients were not statistically significant in the regression.

Setting equal to zero and solving gives the log of optimal capital stock ln K^* (in terms of beds) as:

$$\begin{aligned}\ln\ K^* =& 0.418 + 0.397 \cdot \ln\ A\&E + 0.602 \cdot \ln\ Critical\,Care\,Episodes \\ & + 0.196 \cdot \ln\ Diagnostic\,Services + 0.419 \cdot \ln\ Elective\,Inpatient\,Episodes \\ & - 0.829 \cdot \ln\ \text{Non-Elective}\ Inpatient\,Episodes \\ & - 0.477 \cdot \ln\ Outpatient\,Appointments \\ & + 1.101 \cdot \ln\ Training\,Grade\,Doctor\,Wages \\ & - 3.537 \cdot \ln\ Nursing\,Wages - 0.144 \cdot \ln\ Cost\,of\,Capital \end{aligned} \tag{4.4}$$

Using the expression above, the distribution of optimal capital stock can be compared to observed values. Cost-optimal beds are increasing with all types of activity, except for Non-Elective Inpatient Episodes and Outpatients. Non-Elective Inpatient Episodes require beds, but observed values may appear cost-inefficient due to the requirement to maintain surplus capacity due to the inability of the hospital to turn away cases. In the case of Outpatients, beds are not used. Optimal beds are negatively associated with Nursing Wages, perhaps due to the association between nursing numbers and bed-based care. Where nursing wages are high, hospitals may substitute ward-based care for alternative settings.

When applied to the observed values, optimal bed numbers are slightly higher than observed bed numbers. The distribution of optimal values has a longer tail and higher variance than observed values. The median calculated optimal bed value is 1,070, lower than the median observed value of 741.

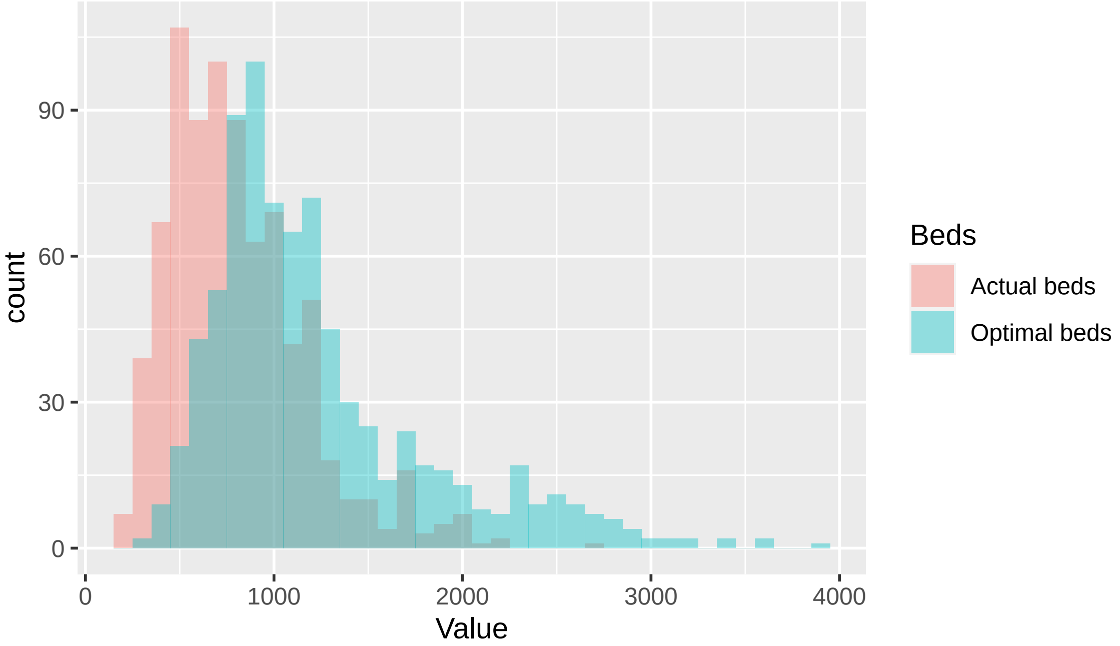

Figure 4.1: Distribution of Calculated Optimal Beds and Observed Beds (Translog). Optimal beds were derived by setting the partial derivative of beds with respect to cost equal to zero and solving. The results suggest that in the period, hospitals were undercapitalised and would lower unit costs with more beds

4.5.1.1 Beds Only Interacted with Output Variables

The estimates presented above include interactions between beds and outputs as well as beds and input prices. Optimal bed numbers would vary according to output levels and also the relative prices of capital and labour. However, input prices that the hospital faces are not in the control of the hospital, and the main relationship of interest is between output and optimal scale. As a robustness check, optimal beds are calculated when interaction terms for beds and input prices are omitted, to see if the optimal bed range calculated above varies.

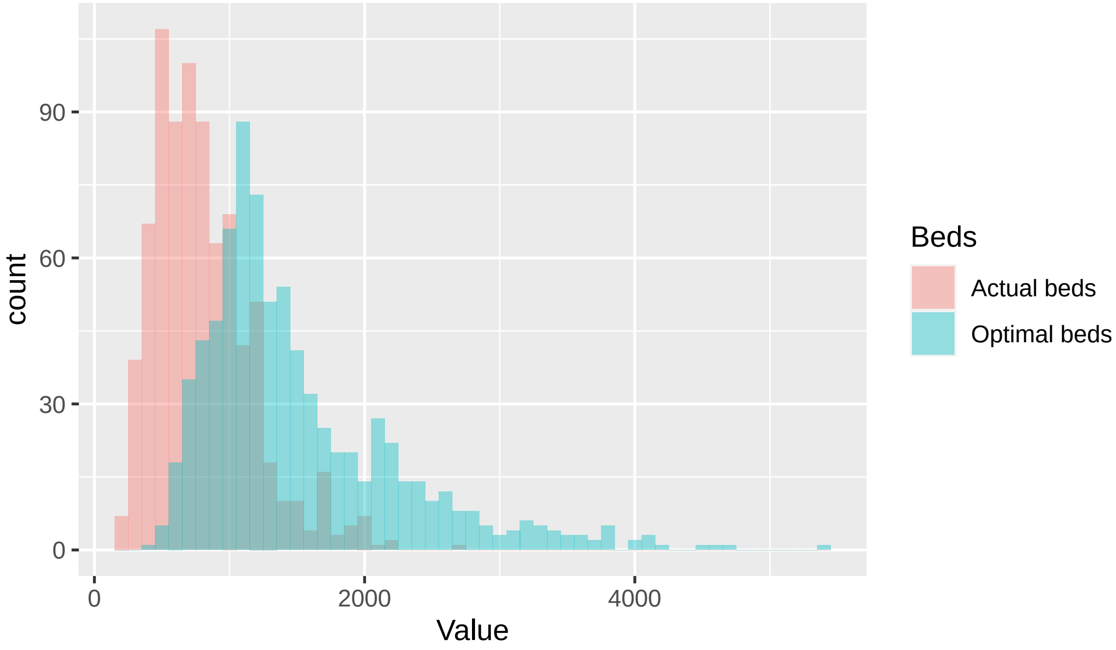

Figure 4.2: Distribution of Calculated Optimal beds and Observed Beds (Translog), interaction terms for output only. See Fig 4.1 for details on derivation and interpretation.

Under both specifications, calculated optimal beds have a similar distribution. Generally, calculated optimal capital levels are slightly higher than actual values, suggesting that hospitals are undercapitalised. Higher capital values would reduce average costs.

4.5.2 Long-run Estimation

Scale economies generated from the short-run estimation are set out in Table 4.1. Retaining beds in the estimation allows us to calculate scale economy values for each observation, factoring in the current number of beds, irrespective of whether this number of beds is optimal for the current output level. As shown in Figure 4.3, scale economies in this case are usually positive. The median short-run scale economy is 1.2, suggesting that a hospital that increased its use of inputs by 1% would see a 1.2% increase in output. Smaller-scale hospitals have higher levels of observed scale economies. The smallest quartile of hospitals (measured by the number of beds) has an average scale economy index of 1.28, with the largest quartile an index of 1.17.

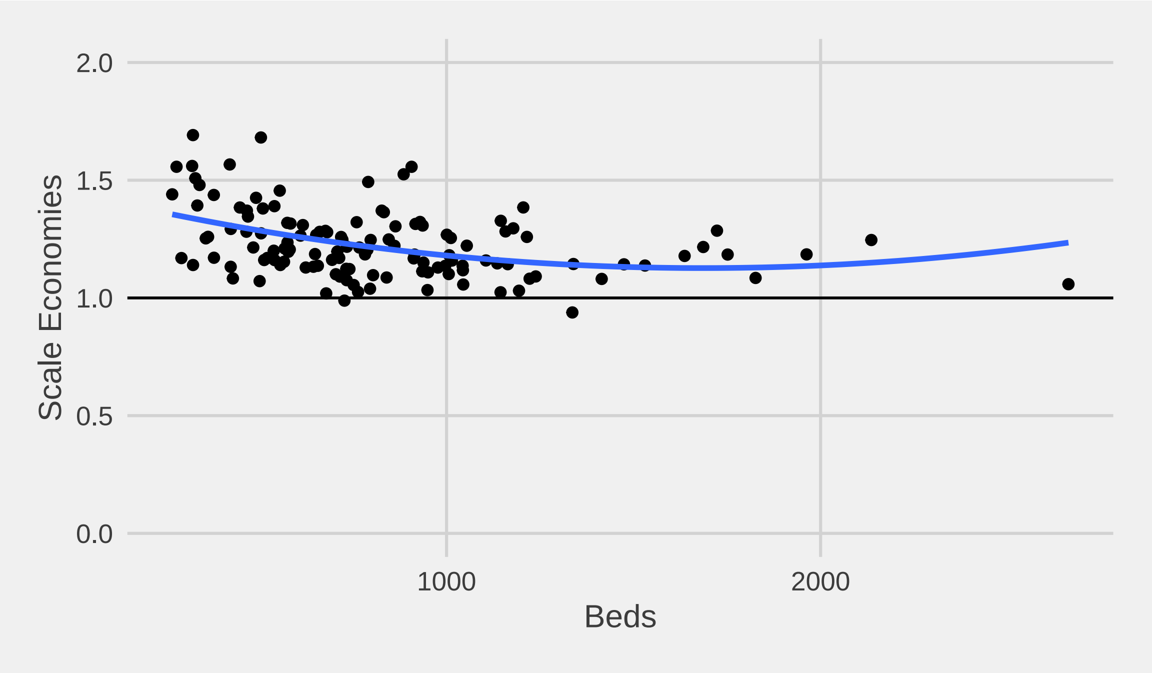

Figure 4.3: Short-run Scale Economies (Translog). Scale economies are derived from the short-run cost function, which includes capital (beds) terms, and are generally higher than direct long-run estimation.

Figure 4.4 shows the long-run scale economy estimates using optimal capital levels, plotted against the current number of beds. Long-run scale economies are, as expected, lower than those calculated for the short-run function. Here the median scale economy index is 1.14. Perhaps most notable in the long-run version is the relationship between scale economies and size. In the direct estimation translog version shown in Section 2.5, scale economies declined slightly with beds, so only smaller hospitals had positive economies of scale. Using the short-run version showed that point scale economies declined with bed numbers, but in the long-run versions, scale economies remained positive at all bed numbers.

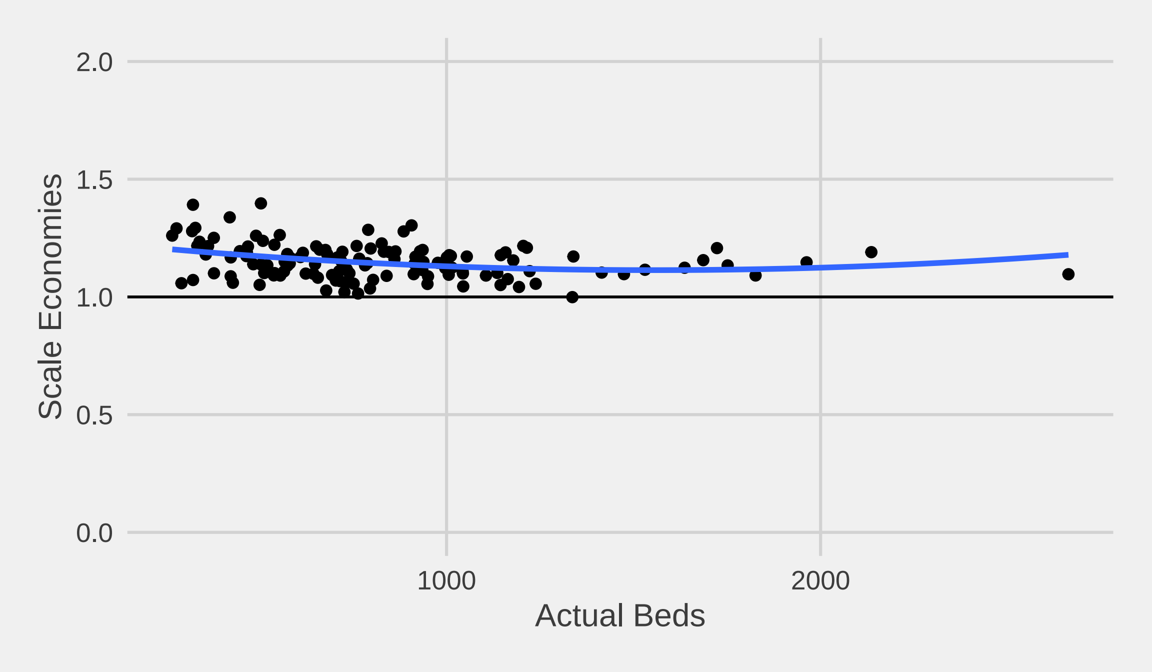

Figure 4.4: Long-run Scale Economies vs Current Beds (Translog). Long-run position evaluated after substituting optimal beds expression into the short-run translog cost function.

Figure 4.5 shows the long-run scale economy calculations plotted against the optimal number of beds. In both this plot and Figure 4.4, calculated scale economies are lower than the short-run calculation in Figure 4.3 but are higher than the directly estimated version in Figure 2.19 from Chapter 2. They are also relatively consistent across the range of actual and optimal beds.

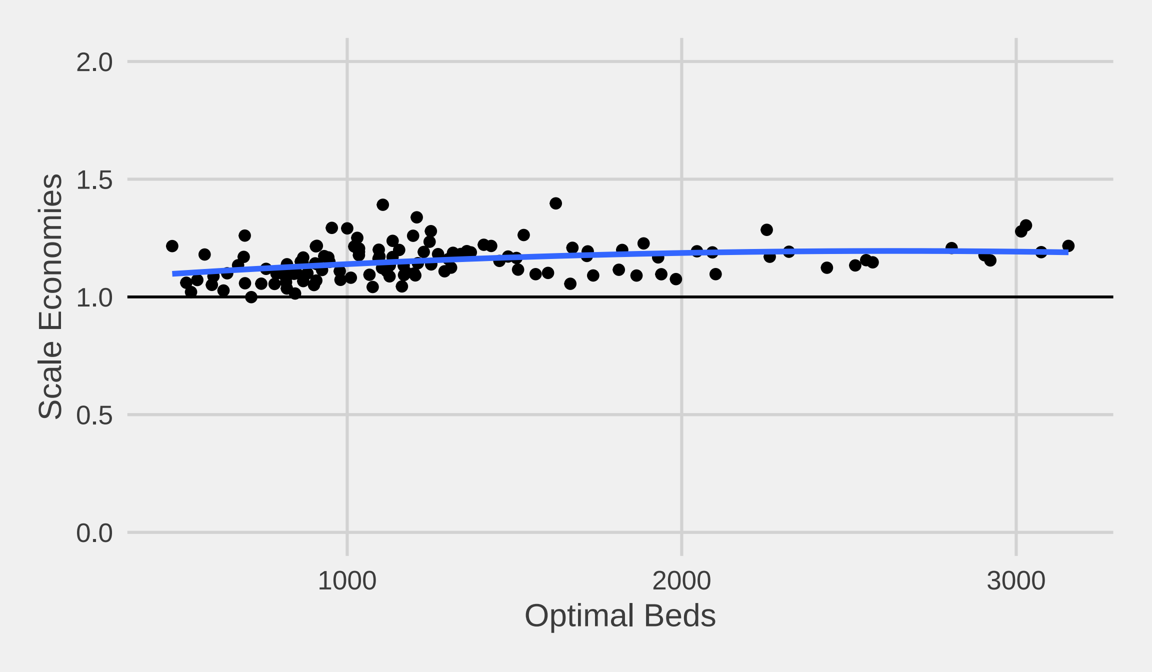

Figure 4.5: Long-run Scale Economies vs Optimal Beds (Translog). Long-run position evaluated after substituting optimal beds expression into the short-run translog cost function.

Output and input coefficients in the directly estimated, short-run and long-run models are shown in Table 4.2, omitting the interaction terms. Compared with direct estimation, output coefficients in the long-run model are higher for A&E Attendances, Critical Care Episodes, Diagnostic Services and Elective Inpatient Episodes and lower for Non-Elective Inpatient care and Outpatient Attendances. Amongst input prices, computing an envelope leads to a lower effect on cost from increases in nursing wages and the cost of capital but a higher effect on cost from increased medical wages.

Table 4.2: Translog Model Coefficients

Term	Direct Estimation	Short-Run Estimation	Long-Run Envelope
Intercept	19.353	19.365	19.395
A&E Attendance	0.076	0.064	0.120
Beds		0.140	
Critical Care	0.078	0.077	0.162
Diagnostic Services	0.048	0.049	0.076
Elective Inpatient	0.301	0.273	0.332
Non-Elective Inpatient	0.269	0.250	0.134
Outpatient	0.144	0.122	0.055
Nursing	0.409	0.462	-0.034
Training Grade Doctor	0.457	0.452	0.606
Capital Rate	0.001	0.000	-0.020

*Interaction terms are omitted

4.5.3 Estimation using a Quadratic Form

As a further check, and to compare how the translog cost function envelope method differs from a quadratic functional form, using the same approach, a long-run quadratic cost function is computed using the envelope approach. Table 4.3 sets out the short-run cost function, including bed numbers:

Table 4.3: Estimation results - Short-run Quadratic Form

Term	Estimate	Std. Error
Intercept	220,749,584.936	(21,640,157.444)***
A&E Attendance	70.578	(153.376)
Critical Care Episode	365.709	(579.409)
Diagnostic Services	0.976	(3.009)
Elective Inpatient Episode	1,044.215	(674.392)
Non-Elective Inpatient Episode	2,042.513	(348.984)***
Outpatient Appointment	114.045	(52.257)*
Training Grade Doctor Wages	1,078.861	(2,123.795)
Nurse Wages	11,920.342	(4,039.651)**
Cost of Capital	3,398,911.814	(3,264,131.632)
0.5 * A&E Attendance * A&E Attendance	-0.001	(0.001)
A&E Attendance * Critical Care Episode	-0.007	(0.003)*
A&E Attendance * Diagnostic Services	0.000	(0.000)*
A&E Attendance * Elective Inpatient Episode	-0.008	(0.004)*
A&E Attendance * Non-Elective Inpatient Episode	0.003	(0.002)
A&E Attendance * Outpatient Appointment	0.000	(0.000)
0.5 * Critical Care Episode * Critical Care Episode	-0.063	(0.008)***
Critical Care Episode * Diagnostic Services	0.000	(0.000)***
Critical Care Episode * Elective Inpatient Episode	-0.004	(0.012)
Critical Care Episode * Non-Elective Inpatient Episode	0.021	(0.006)***
Critical Care Episode * Outpatient Appointment	0.002	(0.001)**
0.5 * Diagnostic Services * Diagnostic Services	0.000	(0.000)***
Diagnostic Services * Elective Inpatient Episode	0.000	(0.000)
Diagnostic Services * Non-Elective Inpatient Episode	0.000	(0.000)
Diagnostic Services * Outpatient Appointment	0.000	(0.000)
0.5 * Elective Inpatient Episode * Elective Inpatient Episode	0.004	(0.016)
Elective Inpatient Episode * Non-Elective Inpatient Episode	0.021	(0.007)**
Elective Inpatient Episode * Outpatient Appointment	-0.002	(0.001)
0.5 * Non-Elective Inpatient Episode * Non-Elective Inpatient Episode	-0.014	(0.006)*
Non-Elective Inpatient Episode * Outpatient Appointment	0.000	(0.001)
0.5 * Outpatient Appointment * Outpatient Appointment	0.000	(0.000)
0.5 * Average Beds * Average Beds	-66.216	(53.370)
Average Beds	122,802.192	(46,743.090)**
Average Beds * A&E Attendance	0.026	(0.186)
Average Beds * Critical Care Episode	1.390	(0.646)*
Average Beds * Diagnostic Services	0.003	(0.004)
Average Beds * Elective Inpatient Episode	0.470	(0.832)
Average Beds * Non-Elective Inpatient Episode	-1.208	(0.427)**
Average Beds * Outpatient Appointment	-0.040	(0.065)

R^2 = 0.9091 adj R^2 = 0.9009 AIC = 28,306 BIC = 28,620 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

Table 4.3: Estimation results - Short-run Quadratic Form

Term	**Estimate**	**Std. Error**
Average Beds * Training Grade Doctor Wages	1.979	(2.597)
Average Beds * Nurse Wages	-10.497	(4.920)*
Average Beds * Cost of Capital	-4,492.919	(3,940.631)
A&E Attendance * Training Grade Doctor Wages	0.027	(0.011)*
Critical Care Episode * Training Grade Doctor Wages	-0.002	(0.042)
Diagnostic Services * Training Grade Doctor Wages	0.000	(0.000)
Elective Inpatient Episode * Training Grade Doctor Wages	0.026	(0.044)
Non-Elective Inpatient Episode * Training Grade Doctor Wages	0.008	(0.024)
Outpatient Appointment * Training Grade Doctor Wages	-0.005	(0.004)
A&E Attendance * Nurse Wages	-0.026	(0.018)
Critical Care Episode * Nurse Wages	0.117	(0.050)*
Diagnostic Services * Nurse Wages	-0.001	(0.000)
Elective Inpatient Episode * Nurse Wages	0.149	(0.073)*
Non-Elective Inpatient Episode * Nurse Wages	0.061	(0.041)
Outpatient Appointment * Nurse Wages	0.005	(0.006)
A&E Attendance * Cost of Capital	-12.114	(14.425)
Critical Care Episode * Cost of Capital	174.014	(52.873)***
Diagnostic Services * Cost of Capital	0.352	(0.255)
Elective Inpatient Episode * Cost of Capital	-59.618	(48.585)
Non-Elective Inpatient Episode * Cost of Capital	27.024	(26.643)
Outpatient Appointment * Cost of Capital	1.441	(4.864)
0.5 * Training Grade Doctor Wages * Training Grade Doctor Wages	-0.117	(0.173)
Nurse Wages * Training Grade Doctor Wages	-0.901	(0.309)**
0.5 * Nurse Wages * Nurse Wages	2.220	(0.728)**
Cost of Capital * Training Grade Doctor Wages	27.255	(158.323)
Cost of Capital * Nurse Wages	141.226	(302.066)
0.5 * Cost of Capital * Cost of Capital	273,561.921	(278,431.378)
Teaching Hospital Organisation	-559,722.581	(9,345,900.311)

$\mathbf{R^2}$ = 0.9091 adj $\mathbf{R^2}$ = 0.9009 AIC = 28,306 BIC = 28,620 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

Optimal beds under a quadratic specification are much higher than observed bed values, in common with the quadratic specification. Optimal beds calculated using the quadratic specification are shown in Figure 4.6. The median calculated optimal bed value is 1911 beds, compared to the median actual bed value of 740. The calculation of optimal beds appears to be sensitive to the form of the cost function specified.

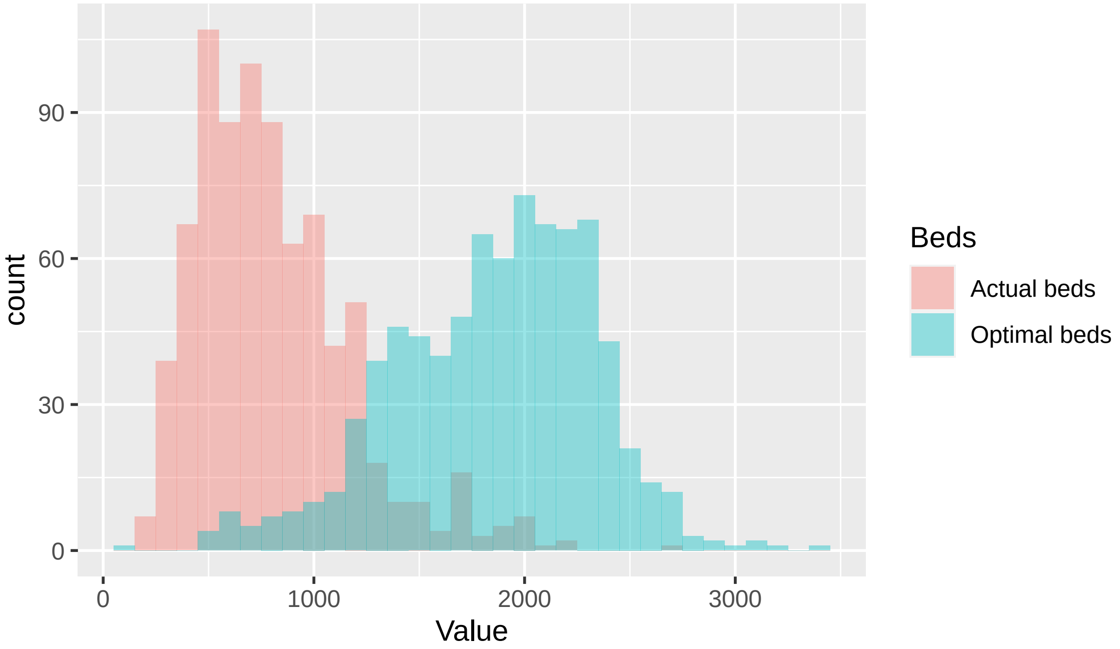

Figure 4.6: Distribution of Calculated Optimal Beds and Observed Beds (Quadratic). See Fig 4.1 for details on derivation and interpretation.

As with the translog function, short-run scale economies (holding current capital levels fixed) are higher than those directly estimated, as seen in Figure 4.7. The translog short-run version has higher scale economies associated with larger hospitals, which was not observed with the quadratic short-run estimation. Median short-run scale economies were 1.185, with no apparent limit to scale economies. Short-run scale economies decline as bed numbers increase but at no point decline to the level of diseconomies of scale.

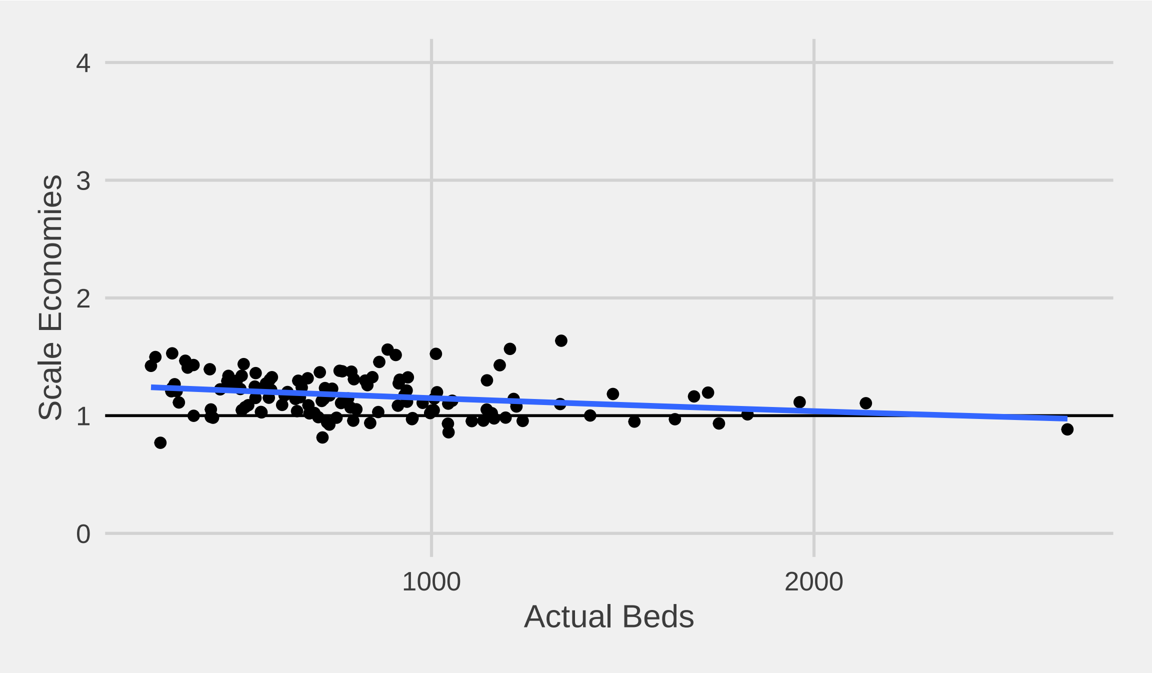

Figure 4.7: Short-run Scale Economies (Quadratic). Scale economies are derived from the short-run cost function, which includes capital (beds) terms, and are generally higher than direct long-run estimation.

The long-run version shows lower scale economies after substituting these optimal bed values into the short-run function. In the long-run most organisations with less than 1,000 beds have positive scale economies (Figure 4.8). When plotting scale economies against optimal beds (Figure 4.9), scale economies are observed to be negative until around 1,500 beds, when they become positive.

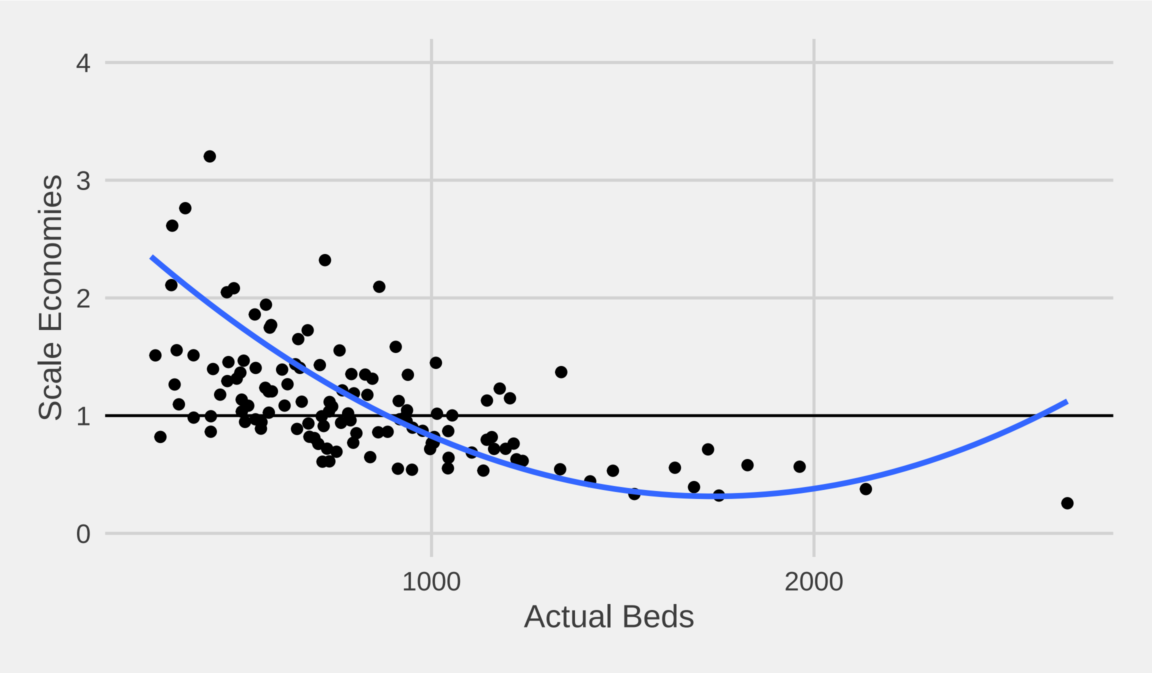

Figure 4.8: Long-run Scale Economies vs Current Beds (Quadratic). Long-run position evaluated after substituting optimal beds expression into the short-run quadratic cost function.

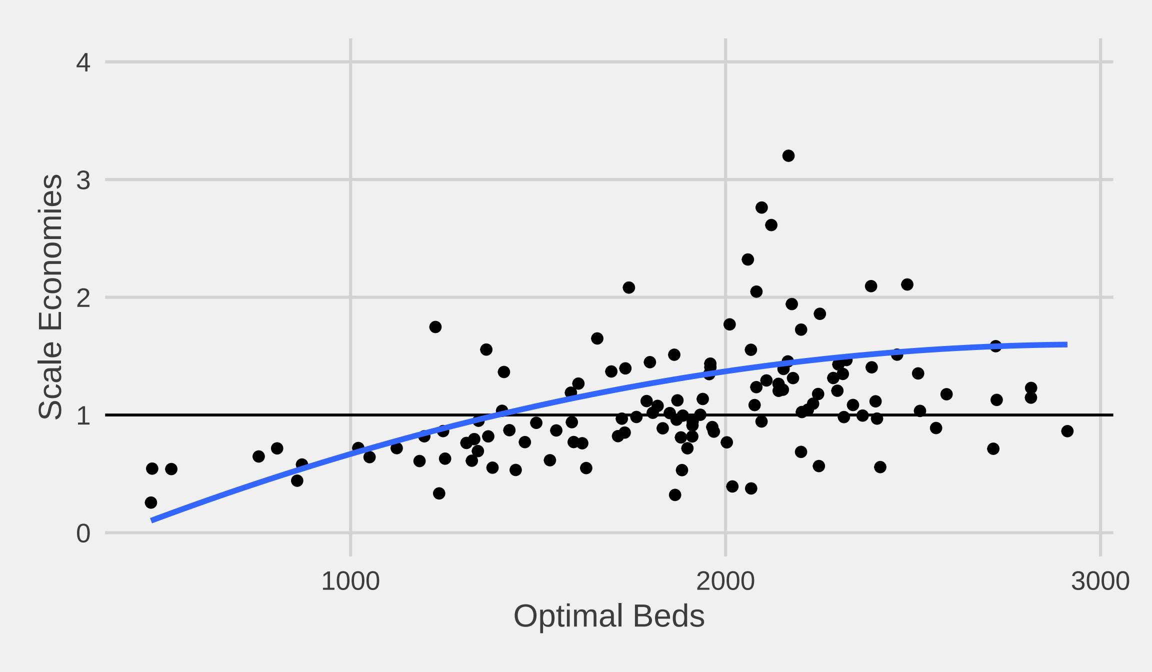

Figure 4.9: Long-run Scale Economies vs Optimal Beds (Quadratic). Long-run position evaluated after substituting optimal beds expression into the short-run quadratic cost function.

4.5.4 Simplified output model

As a further robustness check, a model specification aggregating outputs was estimated to evaluate the sensitivity to changes in output specification. Results are set out in Appendix B, and suggest that calculated optimal capital levels and scale economies are also sensitive to the form and level of aggregation of healthcare outputs, as well as the functional form used for the cost equation.

4.6 Discussion

This chapter used an envelope method to estimate a long-run cost function from short-run functions rather than directly from observed values. Both specifications yielded similar information from the short term regression models, with additional beds being associated with a 0.14% (translog) or £123k (quadratic) increase in total cost. Both functional forms also noted that additional beds were associated with increases in the unit costs of Critical Care Episodes, but were associated with reduced unit costs in non-elective care, and a lower relationship between nursing wages and total cost, perhaps due to a switch to more capital intensive production technologies.

However, results suggest that calculated optimal capital levels are very sensitive to the functional form and output aggregation method. The translog model behaved relatively consistently with the results expected by standard firm theory, with higher scale economies than suggested by direct estimation. However, a quadratic functional form yielded varying results, stemming from implausible calculated optimal capital levels.

More positively, uncertainty in calculated optimal beds is not associated with differences in calculated scale economies. Calculated scale economies appear relatively robust to changes of specification within the same functional form and associated changes in optimal capital calculations. Most variation in computed scale economies in this chapter arises from using different functional forms. Changes in specification within a quadratic or translog functional form do not seem to have as large an effect as changes *between* these forms.

This chapter uses a method of estimating a cost function that deals with the problem of directly estimating scale economies from observed costs, which may not reflect the minimum for each output. This approach ought to correct for this and generate scale economy estimates using the true cost function frontier, improving on studies which use direct estimation. Resulting scale economies and the point at which they no longer hold are higher using this method, as would be inferred from firm theory. The main weaknesses of this approach are the sensitivity of optimal capital levels, which sometimes yield implausible results, particularly for the quadratic specification.

Policymakers can take encouragement from the fact that reported scale economies from direct estimation may underestimate the true size of opportunities from scale economies. However, they may wish to be wary of results from studies that report implausible capital levels or scale economy estimates, as this envelope method appears very sensitive to changes in specification. The quadratic functional form appears more vulnerable to variation in specification and implausible optimal capital results.

Future studies that use this approach to deal with X-inefficiency or non-optimal capital stock should be wary that calculated capital levels are appropriate and are not contingent on the chosen functional form or definition of outputs. Misspecified models may be part of the reason why theoretical papers often identify positive economies of scale (Kristensen et al., 2012; Preyra & Pink, 2006; Vita, 1990), whilst papers looking at post-merger effects may find limited positive effects once mergers have occurred (Gaynor et al., 2012). Future studies could also look at alternate proxy measures for capital, as bed numbers may be vulnerable to the decline in the proportion of hospital healthcare requiring admission. However, at least one prior study using financial statement values rather than beds has found similar difficulties (Cowing & Holtmann, 1983) where optimal capital amounts significantly vary from observed values, suggesting that both capital measures may have issues with sensitivity to specification and returning counterintuitive results. Future studies that use an envelope method should be wary that the calculated long-run cost function is realistic. If optimal capital levels calculated using this method are unrealistic, estimates for ‘optimal’ scale economies may not be realisable. Additional research is needed to establish the sensitivity of such approaches to specification, and whether bed numbers can be used as an appropriate proxy measure for capital.

Chapter 5

Economies of Scope in English Hospitals

5.1 Introduction

Organisations that produce multiple outputs may produce at a lower cost than separate production. These organisations and their production processes are said to exhibit economies of scope. In the case of hospitals, the production of healthcare to treat one health demand may be complementary to the costs of producing other healthcare for different health demands. For example, cost complementarities may arise from the ability to reassign labour to meet fluctuating demand. These complementarities may also arise from the joint use of resources not consumed in production, such as hospital buildings, theatres, or reusable equipment.

In systems where the state plays a direct role in secondary healthcare funding, it is important to know which areas within secondary healthcare exhibit these scope economies. Policymakers can use this knowledge to plan healthcare provision and funding. Discussions on optimal hospital configurations often consider scale effects as sources of potential savings but rarely consider the configuration of services and whether there are financially beneficial scope effects from colocating services. Where scope economies are considered, they are often computed between the care settings, e.g. inpatient and outpatient care, rather than more useful clinical specialty categories.

It is important for policy reasons to consider scope economies across clinical specialties rather than the same units used to measure scale. Policymakers aiming to structure services to maximise scope economies are likely to consider colocating clinical services together, defined in terms of specialties rather than the care setting outputs used in measuring scale. Classifications based on care settings are also vulnerable to casemix differences and associated technical production relationships.

This chapter investigates whether scope economies are present in hospital healthcare, using England's national cost collection data. Data is aggregated into clinical specialty groupings as the most natural way of dividing outputs when evaluating scope economies, rather than retaining an aggregation by care setting as in previous Chapters 2-4. A quadratic cost function is fitted to the national cost collection data, with the largest specialties as outputs and aggregating smaller specialties into a residual category. Having estimated a cost equation, the presence of economies of scope is first tested via cost complementarities, defined as the cross partial derivative of cost with respect to outputs. Following this, the cost difference between separate and joint production is computed using the definition supplied by Baumol et al. (1982).

Results show that 'generalist' specialties, particularly General Surgery, have positive scope economies when combined with other hospital healthcare specialties. Possible reasons for the observed results are discussed, along with suggestions for future research using similar data.

5.2 Previous Literature

Prior studies dealing with scope often do so alongside calculating scale effects. In many cases, this involves using the same output measures used for scale. Using the same units could arise either because the data does not allow other means of aggregation, or because authors wish to consider scale and scope economies using the same output units.

Results from such studies are mixed - some studies find evidence of economies of scope between inpatient admissions and outpatient activity (Sinay & Campbell, 1995; Weaver & Deolalikar, 2004). Other studies find diseconomies of scope between these categories (Grannemann et al., 1986; Wagstaff & López, 1996). Variation in reported results makes it hard to reach a consensus.

One possible explanation for variation in results is the existence of casemix differences. Hospitals may have more or less activity in certain areas based on the mix of healthcare supplied, according to technical relations of production. Certain types of outpatient appointments may be technically related to inpatient care. For example, an elective inpatient surgery admission may be preceded by outpatient appointments to evaluate fitness for surgery or diagnostic imaging, with post-operative outpatient follow-ups taking place after discharge. Accident and Emergency activity is also technically related to Non-Elective Inpatient care. Non-Elective admissions are almost exclusively made via a preceding emergency department attendance. Equally, hospitals may have a casemix which is weighted towards more 'contained' care setting activities, for example, medical care that is carried out in the outpatient setting only, or emergency department activity which has a lower conversion rate to inpatient admission. The degree of scope economies between outpatient, inpatient and emergency care activities may vary according to underlying characteristics in the type of care provided.

Some studies define outputs in more natural units to evaluate scope economies, usually at a clinical specialty level or a grouping of clinical specialties. Prior (1996) uses a DEA approach and finds economies of scope between all combinations of medicine, surgery, gynaecology and paediatric output. Other studies using parametric methods find limited evidence where they are based on the sufficient but not necessary condition of identifying cost complementarities (Vita, 1990), but find more evidence where Baumol's full definition of economies of scope is computed (Baumol et al., 1982, p. 73; Cowing & Holtmann, 1983)

Only one study I am aware of uses UK data to compute economies of scope. Freeman et al. (2020) use English national cost collection data, as is considered here. They find evidence of scope economies between emergency work in a given specialty and elective work in others, and diseconomies of scope for emergency activity in a specialty and elective activity in the same specialty. Rather than estimating an overall cost function, they estimate individual cost functions for groupings of care setting and specialty (defined by the HRG Chapter Hierarchy). Costs are modelled as a linear function of activity in the respective HRG grouping plus another term for activity in other groupings. Freeman et al. can use the dataset to provide weighting within each HRG section and adopt a hierarchical panel structure allowing an analysis of care setting and specialty. However, their definitions of specialties are not exactly congruent with clinical specialisation. For example, HRG Chapter J - "Skin, Breast and Burns" contains many HRGs corresponding to a spell of care primarily under a plastic surgeon, but also breast surgery procedures likely carried out by a general (breast) surgeon, as well as healthcare delivered by a dermatologist (patch tests, phototherapy). Freeman et al. also only consider panels of elective and non-elective inpatient care, perhaps because this HRG-based definition is not amenable to categorising outpatients or emergency department activity. Therefore, they cannot consider scope economies arising from outpatient or emergency department activity. One further issue with the approach Freeman et al. take is that they estimate costs within a care setting rather than at the hospital level. For the left hand side of the model they use total elective (or non-elective) inpatient costs rather than total hospital costs. This approach implicitly accepts the accounting apportionment of organisation-level costs into the two inpatient care setting categories. These apportionments may not be valid if the healthcare production is genuinely joint. For example, apportioned expenditure from building maintenance cannot be logically split between healthcare provided in that building. Treating those costs as apportionable may result in valid scope economies not being detected, where scope economies arise between care settings.

5.3 Data

5.3.1 Clinical Specialities

Medical specialisation has been one of the main techniques to deal with the increasing complexity of healthcare and consequent human capital requirements for physicians. Physicians working in hospital care are specialists in particular areas of medicine, rather than the generalists who predominantly work in primary care clinics. Clinicians working in particular specialties have extended training and particular knowledge in their area of specialisation. This knowledge may not be transferable to other specialties, but it increases the productivity of clinicians working in that field as a form of additional human capital (Marder & Hough, 1983).

Departments based around specialties, or commonly groups of specialties, are the default means of organisation in the UK. Acute hospitals will typically have many medical and surgical specialties coexisting with the hospital. Some specialisms, such as mental health care, are provided by specialist mental health organisations spread throughout the country, and are therefore omitted from this analysis. Equally, some very specialist organisations exist that are national or large regional centres for only one or two medical specialties. These organisations, set out in Appendix A, are also omitted from the analysis.

Certain groups of specialties may benefit from economies of scope. Economies of scope can arise from many sources. For example, economies of scope can arise from the more efficient use of shared resources or physical capital stock, such as surgical theatres, critical care beds, or inpatient wards. Certain specialties may be able to use these common resources where they are colocated in the same organisation and sufficiently physically proximate to each other. More efficient uses of these resources may be possible where one specialty can use resources not currently required by the other, smoothing out fluctuations in demand for healthcare services. The joint production of certain healthcare goods may also require input from several specialties. For example, lower costs in diagnostic specialties such as pathology may reduce time and resource requirements in other specialties using pathology services.

5.3.2 Treatment Function Codes

The data used here is identical to that used in Chapters 2-4. However, in this analysis, the clinical specialty is the more natural activity grouping. In the dataset, medical specialties are differentiated using Treatment Function Codes ('TFC'). Most administrative NHS datasets use these TFCs to denote specialty (NHS Digital, 2022b). The TFC is based on the 'Main Specialty Code', which is determined by which clinician is responsible for the healthcare provided and their specialty as recognised by the relevant UK Royal Medical College. The number and description of TFC codes are set nationally, though individual hospitals decide which code to use for certain activities. Larger hospitals with many specialties and higher degrees of specialisation may use more codes, whereas smaller hospitals may aggregate their activity into higher-level TFC codes. To ensure the data is comparable, and the model is parsimonious, TFC codes are weighted and aggregated into several clusters for analysis, as described below.

5.3.3 Activity Aggregation

Activity is weighted within each specialty and care setting according to the HRG currency value of that activity, before dividing by the average elective inpatient HRG value for all specialties. This approach is analogous to that followed in the simplified model in Appendix B, except that in this version, outputs are not aggregated to a total value but retain an output value for each specialty grouping.

Once weighted, activity groups were aggregated into larger groups according to the following heuristic. If the TFC code had an analogous higher level Main Specialty code, the activity was aggregated into that larger category. For example, TFCs 103 Transplant Surgery and 104 Breast Surgery were aggregated into the higher level category 100 General Surgery. Services which did not have TFC codes were amalgamated into the most relevant Main Specialty grouping. Areas where this was the case included including A&E attendances, critical care episodes, and diagnostics directly ordered by GPs but provided by hospitals. For example, critical care activity was amalgamated into Anaesthetics, as patients on critical care would be under the care of an anaesthesiologist, and GP-ordered diagnostic and pathology services, are amalgamated into the Diagnostic Imaging & Pathology speciality group.

After these steps, progressively smaller specialties were grouped into an 'Other' category until there were eight large speciality groupings, including the 'Other' category. Before aggregation, there were 181 treatment function and service codes, which are now aggregated into eight specialty groups. Smaller specialties are aggregated into an 'other' category so the model is parsimonious, leaving only larger specialties in the model. A complete mapping of how each TFC code is aggregated into a specialty grouping is set out in Appendix C.

Revised weighted activity in each amalgamated specialty group is set out in Table 5.1. The method of aggregation gives us eight distinct specialty groupings for the analysis. Specialties with the highest amount of activity in total include Obstetrics and Gynaecology, General Medicine, Emergency Medicine, Anaesthetics and Critical Care, General Surgery, Orthopaedics and Diagnostics. Each hospital in the dataset recorded activity in these specialties, except for two hospitals which did not offer Obstetrics and Gynaecology services.

Table 5.1: Specialty Activity Levels

Specialty	**Mean Hospital Annual Activity**	**Hospitals Recording Activity**
Other	107,928	141
Obstetrics & Gynaecology	28,296	139
General Medicine	22,531	141
Emergency Medicine (including A&E)	20,326	141
Anaesthetics, Critical Care & Pain Management	20,223	141
General Surgery	18,811	141
Orthopaedics	18,694	141
Diagnostic & Pathology Services	10,376	141

Activity is weighted according to HRG value and setting. Care settings include A&E attendances, Diagnostic Services, Elective Inpatients, Non-Elective Inpatients, Critical Care, and Outpatients. Within each care setting, activity is weighted by the HRG value. The resulting values are weighted by the average cost of each care setting. This value is then divided by the average cost of an elective inpatient episode. The resulting activity value can therefore be considered an 'elective inpatient equivalent' in each specialty. The top 7 specialties are included in the model, with other specialties amalgamated into the 'Other' category

5.3.4 Input Prices

As in Chapters 2-4, average training grade doctor and nursing wages are taken from the NHS's electronic staff record, together with a weighted average cost of capital measure constructed from financial statement information. Though average staff wages may differ by specialty, the average staff wage across the hospital is used. The wage dataset does not allow for identifying wage differentials between specialties. Even if possible, including different wages per specialty would considerably increase the number of terms in the estimation. Consequently, the use of average wages at the hospital level is retained in the estimation.

5.4 Method

5.4.1 Definitions

A discussion of how scope economies apply to multiproduct cost functions can be found in Gravelle & Rees (2004) pp.140-1. Economies of scope exist where the cost of joint production is less than that of separate production. For the two-product case, output values where joint production incurs a lower cost than separate production can be formally stated as follows:

$$C(y_1, y_2, w) < C(0, y_2, w) + C(y_1, 0, w) \tag{5.1}$$

Where this is true over a range of outputs, that range exhibits economies of scope. For higher numbers of outputs, Baumol et al. (1982) derive an expression quantifying economies of scope SC_T as:

$$SC_T = \frac{C(y_T) + C(y_{N-T}) - C(y)}{C(y)} \tag{5.2}$$

Where T is a subset of products evaluated at a particular output vector y. This expression records the difference in costs between separate production $[C(y_T) + C(y_{N-T})]$ and joint production $C(y)$ as a proportion of the joint production cost. Where this expression is greater than zero, the cost of separate production exceeds that of joint production, and there are positive economies of scope. Where this expression is less than zero, joint production costs exceed those of separate production, and there are diseconomies of scope. Where SC_T is zero, the production process is separable, and joint production has no effect on cost.

5.4.2 Cost Complementarities

Cost complementarities exist where the marginal cost of producing one output declines as another input increases. For a twice differentiable cost function, this exists where the following inequality is satisfied:

$$\frac{\partial C(y, w)}{\partial Y_i \partial Y_j} < 0 \tag{5.3}$$

where all $Y_i, Y_j > 0$ and $i \neq j$

Cost complementarities are a sufficient condition for economies of scope (Baumol et al., 1982, p. 75). However, they are not a necessary condition. Economies of scope can arise over certain ranges where the cost complementarity condition is not met. For example, economies of scope may arise where fixed costs are shared but marginal costs are unrelated. The sharing of fixed costs could be sufficient to outweigh the absence of local cost complementarities (Butler, 1995, p. 19; Gorman, 1985). Alternatively, economies of scope may not be present over the entire output range but over a reduced range that encompasses the level at which most organisations produce. Consequently, a test for cost complementaries cannot be a sufficient criterion for ruling out economies of scope.

5.4.3 Estimation of the Cost Function

In this chapter, a quadratic cost function is estimated with outputs aggregated according to clinical specialty rather than care setting, in contrast to previous chapters. Measuring outputs in care settings is helpful for computing economies of scale, where the primary interest lies in overall outputs, and scale economies can be defined as proportionate expansions of the current output mix. However, it is not so useful when considering economies of scope, where the interest lies in how changes in the mix of outputs affects overall costs. Redefining outputs as clinical specialties allows us to analyse the effects of changes in that output mix more usefully. It is more useful to consider specialties as the basis of healthcare goods as they are the most natural ways of thinking about individual services, corresponding to how departments are likely to be structured administratively.

The specific functional form of the estimated function equation (5.4) is identical to the quadratic form used in Chapter 2, except that outputs y_i are aggregated according to specialty groupings rather than the setting in which healthcare is provided.

$$\begin{aligned} C_{ht}(y,w) =& \alpha_0 + \sum_i^m \alpha_i y_{iht} + \sum_k^n \beta_k w_{kht} + \frac{1}{2}\sum_i^m \sum_j^m \delta_{ij} y_{iht} y_{jht} \\ &+ \frac{1}{2}\sum_k^n \sum_l^n \gamma_{kl}\, w_{kht}\, w_{lht} + \sum_i^m \sum_k^n \rho_{ik}\, y_{iht}\, w_{kht} + \epsilon_{ht} \end{aligned} \tag{5.4}$$

Where y_i is the amount of the ith output produced (of m total outputs), w_i is the ith input price (of n inputs), and α_i, β_k, δ_{ij}, $\gamma_{k,l}$, and $\rho_{i,k}$ represent coefficients estimated in the model. Hospital index h and time period t reflect the panel data nature of the dataset. The error term $\epsilon_{ht} = \tau_h + v_{ht}$ comprises a term τ_h representing hospital specific effects and a random noise term v_{it}. For reasons of clarity, time and individual subscripts are omitted from this point on in the chapter.

The quadratic cost function was estimated using a random effects estimator and robust standard errors. Specialties included as outputs are Obstetrics & Gynaecology, General Medicine, Emergency Medicine, Anaesthetics, General Surgery, Orthopaedics, Diagnostic and Pathology Services, and the amalgamated 'Other' category of smaller specialties. As in previous chapters, training grade doctor wages, nursing wages, and the constructed cost of capital are used as input prices, and a dummy variable for teaching hospital status is included. As in previous chapters, the dependent continuous variables are centred to reduce multicollinearity between variables and their cross-products. Centring also means the regression coefficients can be interpreted as the respective effects at the sample mean.

The quadratic form is preferred to the more popular translog to predict costs where specific outputs are set to zero. The translog contains log terms for each output, so an output vector containing zero terms is undefined. Consequently, it cannot be used to predict total cost where certain products are omitted. The quadratic form can predict cost with zero outputs, so it is preferable for calculating economies of scope. The random effects model is used to control for unobserved heterogeneity by taking account of the panel structure. In contrast with the fixed effects model, random effects assumes that the unobserved component τ_h is uncorrelated with other variables in the regression, and that the individual (hospital) specific effect is homoscedastic (Wooldridge, 2002, pp. 247–338).

I investigate the existence of economies of scope in two ways. After estimating a cost function using parametric methods, coefficients on the cross-products of outputs are evaluated. The sign of these coefficients shows us the presence or absence of cost complementarities between products. This approach of testing for cost complementarities is often employed by papers using a translog cost function or as a weaker test for the existence of economies of scope. Following this, scope economies are calculated according to Equation (5.2), using certain clinical specialties as the set T to be tested, with all other outputs included in the set N. Economies of scope are evaluated at the output level of each observation, and the average for each specialty is reported.

5.5 Results

5.5.1 Specialty Activity

Descriptive statistics for the final regression sample are set out in Table 5.2 below, showing summary statistics for the total cost dependent variable and the independent output and price variables, prior to centering.

Table 5.2: Descriptive Statistics

Non-Centred Values	**N = 798**
Anaesthetics	
Mean (SD)	20,223 (17,946)
Median (IQR)	14,089 (8,881, 22,619)
Range	1,696, 108,693
Diagnostic and Pathology Services	
Mean (SD)	10,376 (7,311)
Median (IQR)	8,831 (6,049, 13,286)
Range	632, 99,530
Emergency Medicine	
Mean (SD)	20,326 (10,805)
Median (IQR)	17,715 (13,233, 25,259)
Range	4,228, 84,804
General Medicine	
Mean (SD)	22,531 (12,785)
Median (IQR)	20,808 (14,053, 29,167)
Range	0, 120,049
General Surgery	
Mean (SD)	18,811 (7,275)
Median (IQR)	17,633 (13,522, 23,671)
Range	5,415, 59,031
Obstetrics & Gynaecology	
Mean (SD)	28,296 (14,621)
Median (IQR)	26,949 (18,475, 34,553)
Range	0, 93,639
Orthopaedics	
Mean (SD)	18,694 (7,852)
Median (IQR)	16,821 (13,037, 23,596)
Range	4,104, 46,833
Other Activity	

Table 5.2: Descriptive Statistics

Non-Centred Values	**N = 798**
Mean (SD)	107,928 (62,712)
Median (IQR)	93,367 (60,284, 141,429)
Range	15,614, 379,970
Cost of Capital (%)	
Mean (SD)	4.47 (2.06)
Median (IQR)	3.94 (3.04, 5.36)
Range	0.67, 13.91
Nursing Wages	
Mean (SD)	35,501 (1,930)
Median (IQR)	34,923 (34,260, 35,970)
Range	32,703, 42,253
Training Grade Doctor Wages	
Mean (SD)	47,630 (2,828)
Median (IQR)	47,431 (45,632, 49,543)
Range	40,077, 55,880
Teaching Hospital Organisation	
0	523 (66%)
1	275 (34%)

5.5.2 Cost Function Estimation

Model results are set out in Table 5.3. To solve a computational issue with very small numbers in the inverted variance/covariance matrix, the dependent (cost) variable and all continuous independent variables were divided by 1000, except for the cost of capital variable. The coefficients in the table should be interpreted accordingly.

Table 5.3: Estimation Results - Scope Economies

Term	**Estimate**	**Std. Error**
Intercept	294,323.492	(3,832.179)***
Other	1,165.071	(62.438)***
Obstetrics & Gynaecology	120.103	(209.709)
General Medicine	842.939	(156.718)***
Emergency Medicine Including A&E	911.054	(290.819)**
Anaesthetics, Critical Care & Pain Management	2,785.377	(315.975)***
General Surgery	2,067.344	(458.629)***
Orthopaedics	1,828.620	(356.098)***
Diagnostic and Pathology Services	536.552	(201.970)**
Training Grade Doctor Wages	2,418.323	(453.154)***
Nurse Wages	3,790.022	(1,253.058)**
Cost of Capital	44.547	(649.820)
0.5*Other*Other	-7.873	(2.875)**
Obstetrics & Gynaecology*Other	2.946	(5.296)
0.5*Obstetrics & Gynaecology*Obstetrics & Gynaecology	66.549	(22.930)**
Obstetrics & Gynaecology*Orthopaedics	41.507	(31.346)
General Medicine*Other	4.491	(5.005)
General Medicine*Obstetrics & Gynaecology	-18.339	(14.394)
0.5*General Medicine*General Medicine	-9.645	(16.055)
General Medicine*General Surgery	3.490	(35.274)
General Medicine*Orthopaedics	-37.309	(24.551)
Emergency Medicine Including A&E*Other	0.838	(10.881)
Emergency Medicine Including A&E*Obstetrics & Gynaecology	7.815	(25.218)
Emergency Medicine Including A&E*General Medicine	7.695	(18.046)
0.5*Emergency Medicine Including A&E*Emergency Medicine Including A&E	-22.592	(37.666)
Emergency Medicine Including A&E*General Surgery	12.147	(59.065)
Emergency Medicine Including A&E*Orthopaedics	-20.262	(33.074)
Anaesthetics, Critical Care & Pain Management*Other	23.010	(5.902)***
Anaesthetics, Critical Care & Pain Management*Obstetrics & Gynaecology	-44.203	(12.004)***
Anaesthetics, Critical Care & Pain Management*General Medicine	0.152	(14.802)
Anaesthetics, Critical Care & Pain Management*Emergency Medicine Including A&E	4.561	(29.378)

R^2 = 0.9538 adj R^2 = 0.9488 AIC = 17,586 BIC = 17,961 N = 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19] hospitals

*** - Statistically Significant at 5%**

**** - Statistically Significant at 1%**

***** - Statistically Significant at 0.1%**

Table 5.3: Estimation Results - Scope Economies

Term	**Estimate**	**Std. Error**
0.5*Anaesthetics, Critical Care & Pain Management*Anaesthetics, Critical Care & Pain Management	-70.732	(10.176)***
Anaesthetics, Critical Care & Pain Management*General Surgery	48.264	(41.897)
Anaesthetics, Critical Care & Pain Management*Orthopaedics	-5.222	(33.415)
Anaesthetics, Critical Care & Pain Management*Diagnostic and Pathology Services	-5.617	(15.246)
General Surgery*Other	-10.017	(12.616)
General Surgery*Obstetrics & Gynaecology	-37.636	(46.981)
0.5*General Surgery*General Surgery	62.607	(163.627)
General Surgery*Orthopaedics	-33.907	(85.637)
Orthopaedics*Other	15.300	(11.958)
0.5*Orthopaedics*Orthopaedics	-94.104	(45.502)*
Diagnostic and Pathology Services*Other	-2.605	(7.709)
Diagnostic and Pathology Services*Obstetrics & Gynaecology	16.695	(28.991)
Diagnostic and Pathology Services*General Medicine	22.549	(19.037)
Diagnostic and Pathology Services*Emergency Medicine Including A&E	13.814	(26.147)
Diagnostic and Pathology Services*General Surgery	-24.294	(49.644)
Diagnostic and Pathology Services*Orthopaedics	17.136	(46.683)
0.5*Diagnostic and Pathology Services*Diagnostic and Pathology Services	-17.887	(7.713)*
Other*Training Grade Doctor Wages	-30.208	(18.519)
Obstetrics & Gynaecology*Training Grade Doctor Wages	130.868	(62.662)*
General Medicine*Training Grade Doctor Wages	-38.803	(41.175)
Emergency Medicine Including A&E*Training Grade Doctor Wages	28.856	(84.189)
Anaesthetics, Critical Care & Pain Management*Training Grade Doctor Wages	-47.067	(67.191)
General Surgery*Training Grade Doctor Wages	178.564	(142.868)
Orthopaedics*Training Grade Doctor Wages	-28.973	(89.340)
Diagnostic and Pathology Services*Training Grade Doctor Wages	115.772	(71.654)
Other*Nurse Wages	124.154	(32.715)***
Obstetrics & Gynaecology*Nurse Wages	-223.231	(113.662)*
General Medicine*Nurse Wages	30.452	(87.485)
Emergency Medicine Including A&E*Nurse Wages	-72.618	(122.103)
Anaesthetics, Critical Care & Pain Management*Nurse Wages	99.459	(68.237)
General Surgery*Nurse Wages	-299.676	(281.232)
Orthopaedics*Nurse Wages	250.895	(203.803)
Diagnostic and Pathology Services*Nurse Wages	-100.964	(104.122)
Other*Cost of Capital	-37.195	(16.792)*
Obstetrics & Gynaecology*Cost of Capital	-27.150	(56.784)
General Medicine*Cost of Capital	16.607	(52.540)
Emergency Medicine Including A&E*Cost of Capital	-196.704	(81.850)*

R^2 = 0.9538 adj R^2 = 0.9488 AIC = 17,586 BIC = 17,961 N = 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19] hospitals

*** - Statistically Significant at 5%**

**** - Statistically Significant at 1%**

***** - Statistically Significant at 0.1%**

Table 5.3: Estimation Results - Scope Economies

Term	Estimate	Std. Error
Anaesthetics, Critical Care & Pain Management*Cost of Capital	248.437	(65.738)***
General Surgery*Cost of Capital	-121.525	(146.403)
Orthopaedics*Cost of Capital	213.906	(134.866)
Diagnostic and Pathology Services*Cost of Capital	143.385	(72.427)*
0.5*Training Grade Doctor Wages*Training Grade Doctor Wages	122.411	(209.510)
Nurse Wages*Training Grade Doctor Wages	-892.272	(371.245)*
0.5*Nurse Wages*Nurse Wages	2,611.275	(809.077)**
Cost of Capital*Training Grade Doctor Wages	60.303	(171.173)
Cost of Capital*Nurse Wages	835.107	(403.547)*
0.5*Cost of Capital*Cost of Capital	633.154	(272.281)*
Teaching Hospital Organisation	-4,864.575	(5,751.117)
London-based Organisation	34,305.599	(10,549.015)**

R^2 = 0.9538 adj R^2 = 0.9488 AIC = 17,586 BIC = 17,961 N = 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19] hospitals

* - Statistically Significant at 5%

** - Statistically Significant at 1%

*** - Statistically Significant at 0.1%

Significant coefficients in these results include the intercept term, which at £294,323,492 is comparable to estimates in previous chapters. The dummy for London-based organisations is also significant and estimated at an additional £34,305,599 cost associated with a location in London. The non-interacted output coefficients are all positive and statistically significant except for Obstetrics and Gynaecology. Coefficients on non-interacted input prices are also all positive and statistically significant except for the cost of capital variable. Interaction terms were generally not statistically significant, but terms involving self-cross products were often statistically significant. The self-cross-product of Obstetrics and Gynaecology (£67) is statistically significant and positive, suggesting that costs increase with output for this specialty. The self-cross-product of the ‘Other’ category (£-8) suggests that marginal costs in this category decline with additional output. Similar observations were noted in orthopaedics (£-94), Diagnostic and Pathology Services (£-18), and Anaesthetic, Critical Care and Pain Management (£-71), suggesting weak evidence for scale economies in these specialties.

Some price-output interactions were found to be statistically significant. A £1 increase in Training Grade Doctor wages was associated with a £131 increase in Obstetrics and Gynaecology Costs per activity unit. Marginal costs in Obstetrics and Gynaecology decline by £223 per activity unit with a £1 increase in nursing wages. This finding may be a seniority effect caused by midwives and other nursing staff in this group being in lower pay bands on average than other specialties. Higher activity in Obstetrics & Gynaecology may be therefore associated with lower nursing wages and, therefore, a lower total cost. Higher nursing wages are associated with higher marginal costs in the 'Other' specialty grouping, adding £124 to each activity unit for each £1 increase in nurse wages. A 1% increase in the cost of capital is associated with lower marginal costs of the 'other' specialty grouping by £-37, lower marginal costs in Emergency Medicine (£-197), higher marginal costs for Diagnostic and Pathology Services (£143) and Anaesthetics, Critical Care and Pain Management (£248).

Several input-interaction terms were statistically significant. The interaction of nurse wages and training grade doctor wages is negative and statistically significant (£-892), suggesting a degree of substitutability between medical and nursing labour. Higher paid or higher graded staff in one staff category reduces the marginal impact on cost from raises in the other staff group. The marginal impact of nursing wages on total cost unsurprisingly increases as nursing wages increase (£2,611). A similar phenomenon arises from increases in the cost of capital (£633). Finally, increases in nurse wages are associated with increased marginal effects of the cost of capital on total cost. Similar effects were observed in earlier chapters and hypothesised as the result of nursing labour being complementary to additional capital requirements in the form of wards.

5.5.3 Cost Complementarities

Positive and statistically significant cost complementarities were observed between the Anaesthetics and Obstetrics/Gynaecology groupings. At the sample means, an additional 'inpatient equivalent' unit of activity in one specialty reduced the cost of an activity in the other by £44. This reduction could result from scope efficiencies for activities in surgery or obstetric pain relief. However, it is notable that the coefficient for the cost complementarity between Anaesthetics and General Surgery is not statistically significant and has a positive sign suggesting diseconomies of scope. The only other statistically significant cost complementarity was observed between Anaesthetics and the 'Other' category, which comprised the aggregated smaller specialty groups. The coefficient has a positive sign, suggesting diseconomies of scope, with an additional activity unit in Anaesthetics increasing the cost of an additional activity unit in the 'Other' category by £12, and vice versa. No other statistically significant cost complementarity effects were observed. Using cost complementarities as a measure of scope economies suggests a limited effect for the specialty groups tested.

5.5.4 Scope Economy Indices

Scale economy calculations were computed according to Equation (5.2). The distribution of the calculated scope economy index at each data point is shown in Figure 5.1. The scope economy index was highest for General Surgery, with a median value of 0.14. This index value implies that joint production is 14% less expensive than separate production over the range of outputs considered in the study population. Other specialty groupings with positive median scope economies included Obstetrics/Gynaecology and General Medicine. Obstetrics/Gynaecology had a median scope economy index of 0.06, with joint production being 6% less expensive than separate production. 14% of observations have negative scope economies, leaving 86% with positive scope economies. The scope index for General Medicine had a median value of 0.04, with joint production being 4% less expensive than separate production. 83% of observations showed positive General Medicine scope economies, and 17% of observations had negative General Medicine scope economies.

Other specialty groups recorded inconclusive or negative scope economies. The Diagnostic and Pathology Services specialty grouping had a median scope index of -0.01, indicating that separate production was 1% less expensive than joint production. However, the distribution of scope economies lies on either side of the zero dividing line, with 34% of values recording positive scope economies and 66% negative values. The median scope economy index for Orthopaedics was -0.04, with separate production being 4% less expensive than joint production. 73% of observations had negative values, and 27% were positive. The median Emergency Medicine and A&E scope index was -0.03, suggesting that separate production was 3% cheaper than joint production, with 75% of observations showing negative and 25% positive scope economies. Most observations also showed negative scope indexes for Anaesthetics and Critical Care, where the median scope economy index was -0.09, 87% of observations with negative scope economies, and 13% positive scale economies. The 'Other' category had a median scope index of -0.21, with 94% of observations having negative scope economies and 6% positive scope economies.

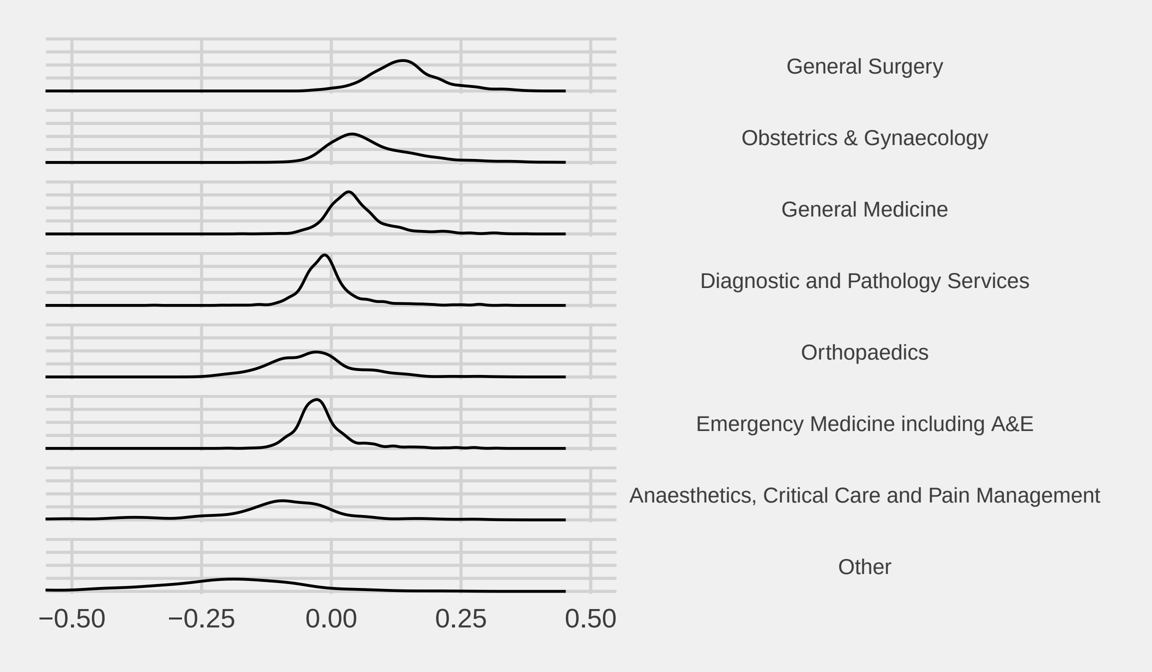

Figure 5.1: Distribution of Scope Economy Index for eight specialty groupings, including the residual category 'other'. Scope economies are calculated with reference to all activity other than the specialty group in question. Index values > 0 suggest positive scope economies at the current output level. General Surgery, Obstetrics & Gynaecology and General Medicine have positive economies of scope. Diagnostic and Pathology Services, Emergency Medicine, Orthopaedics, Anaesthetics, and the residual 'Other' category have zero or negative scope economies.

5.6 Discussion

This chapter shows that certain specialties are associated with stronger scope economies when combined with others. General Surgery has strong scope effects, with joint production being 14% less expensive than separate production for the median hospital. Scope effects reflect the variety of surgical work undertaken by General Surgeons and their work being complementary to other specialties where surgery is required. In the UK, subspecialties of General Surgery include breast surgery, colorectal surgery, upper gastrointestinal surgery, endocrine surgery, and transplant surgery (Royal College of Surgeons, 2022). Each of these subspecialties may overlap with activity in other specialties. Hospitals with larger General Surgery services may benefit from greater in-house capabilities and a higher degree of specialisation within General Surgery. Increased specialisation within General Surgery could positively affect costs in other specialties. Equally, theatre resources used by general surgeons could also be made available to other surgical specialties, as could postoperative beds. Sharing these resources could yield savings, both from avoiding multiple purchases, but also from using these resources more intensely, avoiding unproductive time where resources are not used. Interestingly, there is little evidence for scope economies in Orthopaedics, suggesting that the sources of these scope economies may be more prevalent in General Surgery than other surgical specialties.

The other specialty grouping showing positive scope economies was General Medicine. General Medicine also has technical reasons why co-production may lower costs. Medical staff working in General Medicine often have a dual specialty, so they may be able to substitute for labour in other departments. Wards used in General Medicine may be able to accommodate patients as an alternative to other medical specialties. Taking these two results together, it is apparent that specialties which retain a 'generalist' outlook or historical focus yield scope economies. These specialties, their subspecialties and closely related specialties ought to be colocated to realise scope economies.

Obstetrics and Gynaecology also showed positive scope economies. Cost complementarities were observed between this specialty grouping and Anaesthetics. Obstetric activity may share resources with Anaesthetic activity, for example, in providing pain relief and neonatal critical care. Gynaecology activity also uses Anaesthetic and Critical Care resources like other surgical specialties. Consequently, shareable resources between Obstetrics/Gynaecology and Anaesthetics may be higher than for other specialties.

Limited scope economies were found for Diagnostic and Pathology Services. This result suggests that consolidating pathology services into larger providers would not lose economies of scope from lack of colocation. Transfer of pathology samples can be made relatively easily at a low cost, and specialised pathology providers can benefit from scale economies where they provide services for several hospitals.

A&E displayed negative scale economies, with increased A&E activity effectively raising costs in other specialties. This finding could be attributable to imperfect activity weighting and higher resource requirements for non-elective admissions through A&E compared to elective and planned activity. Higher A&E activity could raise costs elsewhere because of casemix changes towards more resource-intensive cases. Alternatively, higher A&E activity could raise costs elsewhere by increasing the amount of 'surplus' capacity and resources in other specialties that would be required to deal with the inherently variable nature of emergency care.

This chapter contributes to the limited literature on scope economies, and particularly improves on prior studies, which often use the same output definitions for scale and scope, typically based on the method of admission or care setting. Knowing that there are scope effects between elective and emergency care may be useful, but it is more useful to consider output based on clinical specialty, as this is the typical basis for organising services and constructing hospital departments.

The conclusions drawn from this approach depend heavily on the aggregation of activity and clinical activities into the specialty groupings. In some cases, an alternate grouping would be as justifiable. For example, paediatric ENT activity could justifiably be recorded under paediatrics or ENT. Surgical resection of tumours could likewise be categorised under surgery or an oncology grouping. Different aggregations and assignments of activity could yield different results, making the comparison to other studies complex where studies define specialties in different ways. Comparisons to other studies could be affected by differences in how hospitals record and attribute activity. It is important to be careful in generalising results from one health system to another.

Evaluating the cost effects of colocation does not consider technical relationships between certain specialties and departments. The analysis finds limited evidence of scope economies between Anaesthetics (including Critical Care) and other specialty groupings. However, anaesthetic activity and Critical Care are technical requirements for many surgical specialties. Colocation would be technically required even where scope economies are negligible, as surgical activity will require anaesthetic input, and many specialties may require patients to spend time in critical care wards. Therefore, technical or clinical requirements may require colocation even if there are limited financial benefits.

Policymakers working in secondary care need to know which specialties are best colocated together for clinical and economic reasons. Having General Surgery, General Medicine, and Obstetrics/Gynaecology alongside other clinical specialties in a hospital could be expected to yield positive spillover benefits. Whilst there are clinical reasons why specialties like these need to be spread widely and available locally, it is also true that there appear to be economic benefits from doing so. Attempts to concentrate such activity in fewer hospitals could increase average costs overall if scope effects outweigh scale benefits. These specialties ought to be a part of the minimum services that hospitals offer - as in most cases they already are.

Researchers working in this area can note that the analysis improves upon studies that test for the existence of cost complementarities as their sole means of identifying scope economies. Testing for cost complementarities cannot identify cases where scope economies arise from sharing fixed resources. The high value of the intercept term in this study suggests that sharing fixed factors of production could be an important source of scope economies for hospitals. If true, the existence of scope economies would be missed by studies using cost complementarities only.

This analysis considers an aggregated view of specialties and comparators. Future research could expand on this by more detailed and smaller scale comparison. Such analysis could be carried out via pairways comparison of individual specialties or by increasing the number of specialties included in the model. Care needs to be taken so that the specification does not become unwieldy. To facilitate comparison, studies should also publish the means of aggregation and their definitions of activity. Future research could also investigate the source of scope economies. The cost function estimated above showed limited cost complementarities but positive scope economies and a high level of fixed costs as represented by the intercept term. These phenomena may be reconciled if the sources of scope economies lie in the sharing of resources and fixed factors of production rather than marginal effects where the expansion of one specialty lowers costs in another. Further research is needed to evaluate this hypothesis and develop our understanding of hospital economies of scope.

Chapter 6

Conclusion: Size, Scope and Hospital Costs

6.1 Aims

This thesis estimates hospital scale and scope economies from observed outputs and input prices, using several parametric techniques and functional forms of the cost function. These investigations provide valuable results for our understanding of scale economies in hospital healthcare. Estimates of scope economies are provided for major specialty groups, showing which clinical specialties are most likely to benefit from scope economies when co-located with all other specialty groups. These estimations provide helpful information for researchers, policymakers and those managing secondary care in service delivery terms who are interested in the relationship between size and cost. This thesis contributes to research in the area by updating previous estimates that are in many cases dated. It also provides empirical estimates from a country (England) where these questions are not often studied. Additional and potentially more useful methodological results are obtained using multiple functional forms and estimation methods. This approach contrasts with other studies that typically use one method. More information adds to these results' potential usefulness in decision making. The results of these essays inform our understanding of how hospital cost and size are related, deepening the evidence base and expanding it to new areas.

6.2 Findings

6.2.1 Differences between Quadratic and Translog Functional Forms

In Chapter 2 and Chapter 3, scale economy estimates are derived from administrative data sources. Data sources included activity and cost data from NHS England's National Cost Collection Exercise, wage data from NHS Digital's Workforce Statistics, and a capital cost constructed from NHS hospital financial statements. These data sources are used to estimate a long-run cost function directly from observed values. Point scale economy estimates are computed for each observation as the inverse of the sum of cost-output elasticities - representing the percentage increase in output obtained by a 1% increase in costs. These point estimates were plotted alongside measures of size, chiefly the number of hospital beds. The relationship between the point scale economy estimates and size was evaluated for evidence of any limits to scale economies. In Chapter 2, a translog functional form is used for the cost function, contrasting with a quadratic functional form in Chapter 3.

Both translog and quadratic functional forms found positive economies of scale up to 1,200 beds. Beyond this, the results were less consistent. Scale economies estimated using a translog function declined after this point to a level best characterised as constant economies of scale. Scale economy estimates using a quadratic specification were more varied and did not decline with size, with most observations remaining positive at all sizes. As identical data is used for both, this result may be due to the technical properties of the functional form. Both functional forms offered a similar degree of consistency with theoretical principles, with the translog being marginally more consistent with respect to the convexity of certain variables. Few previous studies use multiple functional forms, so this thesis suggests that differences in results between translog and quadratic forms may result from the properties of the specification. As such, caution is needed when interpreting studies that use only one form. In this sample, positive economies were observed amongst hospitals with up to 1,200 beds. Beyond that size, scale economies are less certain. It would be interesting to see further research in other countries on large data sets to see how similar results compare.

6.2.2 Differences in High Cost Areas

In both quadratic and translog functional forms, scale economies were lower in the higher wage areas comprising London and surrounding areas. This qualitative difference suggests that hospital care supply policy may differ in high-cost areas, as increases in output cost proportionately more than in other areas. Whilst hospital healthcare supply needs to be proximate in most instances, results suggest that hospitals in these regions may not benefit from scale to the same extent as elsewhere. This finding may need to be addressed by healthcare commissioners and funding allocations. It may be the case that hospitals in higher wage areas may require different production technologies and a different capital-labour mix. This observation is another potential area for further investigation.

6.2.3 Direct Estimation of Long-run Cost compared to Enveloping Short-run Cost Functions

Chapter 4 compares the directly estimated long-run cost functions with those derived from the envelope of short-run cost functions. Both translog and quadratic cost functions are estimated from the same data sources. However, in this chapter a short-run cost function was estimated by including a capital (beds) term and appropriate interactions in the estimation. Short-run scale economies were estimated using the same method as in Chapters 2-3, excepting the different specification. To allow for the possibility that observed capital values were not optimal at the time of observation, optimal values were computed by taking the partial derivative of the cost function with respect to beds, setting the expression equal to zero and solving. This expression for optimal beds was substituted back into the cost function to yield an optimal long-run cost function which is functionally the envelope of the short-run cost curves. This revised long-run function was then used to compute long-run scale economies as in the directly estimated case.

Scale economies computed using this indirect method were higher than directly estimated calculations, as was previously suggested by authors critiquing the direct estimation method. However, estimated optimal capital levels computed by this method seem to be much higher than current levels. The calculated long-run quadratic cost function seemed vulnerable to implausible optimal capital levels. As a result of these implausible optimal capital levels, estimated long-run scale economies using a quadratic specification actually increased with size, with no discernible limit. The translog specification returned more plausible optimal capital levels, though in most cases below observed bed numbers. Such variability observed may be a function of the relatively poor proxies for capital available - bed numbers may be increasingly less appropriate proxies where stay lengths decline and ambulatory care increases. However, differences between calculated optimal capital levels between the translog and quadratic specifications suggest that this method may be very sensitive to the choice of functional form. Where calculated optimal capital levels are not accurate, there is a risk that the calculated long-run cost function is based on unachievable points, vulnerable to specification or choice of functional form. Though this method has the attraction of dealing with the issue of observed values not being on the optimal cost frontier, the method for resolving this issue could result in an estimated cost function with lower long-run costs that are, in fact, achievable. This observation is an interesting finding, informing future research directions in methodological and applied terms, and future research using improved measures of capital may be a useful next step.

6.2.4 Scope Economies

Chapter 5 sets out results from a model making an initial investigation of scope economies. A quadratic cost function is estimated, with outputs reconfigured as clinical specialties rather than as weighted case numbers in particular settings. This output specification is a more useful aggregation of healthcare outputs, as groups of clinical specialties are typically the basis for hospital departments. The existence and extent of scope economies are investigated firstly by tests for cost complementarities in the results. As positive cost complementarities are a necessary but not sufficient test for scope economies, The cost difference between joint and separate production is directly calculated for each specialty group and observation, as generated by the estimated cost function.

Clinical specialties which are more ‘generalist’ or are closely related to activity in other specialties, such as General Surgery or General Medicine, showed much larger economies of scope than other specialties when compared against all other activities. These specialties often assist or provide associated care for other clinical specialties. Consequently, having a reasonably sized General Surgery or General Medicine department will lower costs elsewhere, perhaps by allowing general surgeons or general internal medicine doctors to specialise more *within* these specialties. Alternatively, clinical labour in these specialties may be more readily substituted for other specialties to flexibly respond to demand.

Obstetrics and Gynaecology also showed positive scope economies, with the model results showing particular cost complementarities with a grouping of anaesthetics and critical care. Diseconomies of scope were found for Accident and Emergency activity compared to other specialty groupings, perhaps because a high level of Accident and Emergency activity would require other clinical specialties to maintain sufficient capacity for Non-Elective admissions. The additional capacity required for non-elective care in other specialties may not be fully used, increasing average costs in hospitals with considerable Accident and Emergency activity relative to those with more planned care. Hospitals providing relatively more planned care would not need to account for peaks in demand for non-elective care and would not need surplus capacity standing idle in periods of low non-elective care demand.

6.3 Policy Implications

6.3.1 Scale Economies (Small & Medium-sized Hospitals)

Both the directly estimated translog and quadratic specifications showed small but positive economies of scale up to around 1,200 beds, higher than previous studies. Policymakers can take some assurance that at (pre-pandemic) output levels, the majority of hospitals may lower average costs by expanding output. The directly estimated translog cost function had the lowest median scale economy index at 1.04, with higher estimates for hospitals smaller than the median hospital. Using this estimate, the median hospital would see output increase by 1.04% for a 1% increase in costs. The directly estimated quadratic functional form returned a comparable median scale economy estimate at 1.03. Studies using an envelope returned higher values, up to a 1.5% increase in output for a 1% increase in costs using a translog functional form. However, results using an envelope method are vulnerable to weaknesses in the available proxy measures for capital and may consequently be sensitive to changes in specification. This observation was particularly evident in the case of the quadratic functional form, which returned implausible optimal capital results.

There are multiple policy options to realise gains from scale. In an environment where overall hospital healthcare demand is growing, policymakers can influence where the growth takes place, directing expansion at smaller hospitals with larger scale economies, growing activity and lowering unit costs at these hospitals. The other method of realising gains would be an element of consolidation, closing some of the smaller hospitals to provide healthcare at scale in other hospitals. Consolidation perhaps makes more sense in an environment where healthcare demand or funding is relatively static. Where growth in the sector is limited, consolidation would be required to realise gains.

Consolidation of the sector is a complicated area. Given the magnitude of scale economies, it is likely that short-term consolidation costs would outweigh the gains so that any benefits may be limited to the long-run. However, even if consolidation were desired in the longer-term, overall cost considerations are not the only concern. There will often be a clinical need for access to proximate care for health emergencies, which requires the maintenance of smaller hospitals in rural areas, even where unit costs are higher than in larger urban centres. Though this study does not consider equity explicitly, any consolidation also has equity considerations. Closing smaller hospitals may disadvantage specific demographics in those areas at the expense of those living near larger hospitals. Cost would not, of course, be the only policy consideration in any change to healthcare supply.

6.3.2 Scale Economies (Larger Hospitals)

The evidence for larger hospitals, defined as those with more than 1,200 beds, is more mixed. There are fewer hospitals of this size, so unsurprisingly, differing functional forms and methods return different scale economy estimates. Directly estimated translog scale economies were broadly constant, whilst those obtained using an envelope method were slightly positive. A quadratic specification showed positive scale economies when directly estimated but diseconomies of scale using an envelope method, likely due to infeasible optimal calculated capital levels. Policymakers should therefore be cautious when dealing with larger hospitals and avoid assuming that widespread scale economies continue without limit.

6.3.3 Scale Economies (Geographical Differences)

This thesis also shows that scale economies are lower in London and surrounding areas due to higher wage rates. Policymakers could look to expand the healthcare supply in hospitals close to but outside the areas where higher wage rates apply. Healthcare supply for certain types of care needs to be proximate for clinical reasons, and there are geographic limits to how far patients can be expected to travel for care. However, where possible, policymakers could realise additional benefits from expanding activity at hospitals close to the higher wage area. Policymakers may wish to bear geographic differences in mind, such that hospitals in higher wage areas may not benefit from scale to the same extent as elsewhere.

6.3.4 Scope Economies

Chapter 5 showed that General Surgery saw the highest scope economy measure for the specialty groupings considered. Policymakers wanting to take advantage of this result could ensure that each hospital offering surgery has sufficient General Surgery services, as these are likely to lower costs in other specialties. General Surgery, along with General Medicine and Obstetrics/Gynaecology, should be a core part of each hospital's supply of secondary healthcare. Conversely, the lack of scope economies for diagnostic services suggests that these services may be more suited to amalgamation, especially if activity within these services demonstrates scale effects. Negative scope economies associated with Accident and Emergency activity might suggest that hospitals providing A&E services be provided with additional funding, accepting that they have negative spillover effects on other specialties.

6.4 Research Implications

6.4.1 Contribution to Knowledge

This thesis provides updated estimates for scale economies in the supply of hospital healthcare, using data from England. It updates previous studies that are now quite dated and demonstrates the existence of small but positive economies of scale up to 1,200 beds. Few hospitals in the dataset were larger than this, and after this point, different methods and functional forms of the estimation produced scale economy estimates ranging from constant to positive returns to scale. This study is the first for some time to estimate a particular limit to economies of scale using English data. It also contrasts several parametric methods employed in the literature, showing variation in different techniques using the same dataset. Few previous studies contrast different methods using the same dataset, so this thesis can demonstrate differences in results that may result from differences in functional form and estimation method.

This study also uses English data to estimate scope economies. Few prior studies investigate scope economies, and even fewer have used clinical specialties as a unit of output. As hospital departments are structured around clusters of clinical specialties, this would be a natural unit of production for scope economies. Many studies that estimate both scale and scope economies will use output definitions appropriate for measuring scale (admitted care cases, ambulatory care cases, other activities) but not appropriate for scope. Policymakers would define ‘departments’ or sub-units of production using clinical specialties rather than the type of admission. The results in Chapter 5 show that General Surgery benefits most from scope economies, with General Medicine and Obstetrics/Gynaecology also showing positive but lesser scope economies. I am unaware of any prior study using UK data which looks at scope economies between specialties, so these findings appear novel to the literature.

6.4.2 Methodological Implications

There were observed differences between translog and quadratic results, especially amongst the range of larger hospitals where fewer observations are available. Studies that use only one type of functional form should at least use an alternate form as a robustness check. The requirement for sensitivity checks is greater for studies which attempt to measure limits to scale economies. Calculated limits to scale economies are more likely to be at the larger end of the size distribution, so the estimation of limits may be more vulnerable to differences in the functional form chosen.

This thesis has also shown differences between estimated scale economies between a directly estimated cost function (Chapter 2 and Chapter 3) and an alternate method using an envelope of short-run cost functions. However, using an envelope constructed from 'optimal' beds yielded implausible capital values, particularly in the case of the quadratic functional form. Studies which use the envelope method may be misestimating scale economies where they cannot calculate optimal capital levels. Suppose the number of beds is not an effective proxy measure for capital. In that case, the envelope method may overstate the extent to which average cost curves are lowered at 'optimal' capital levels. In seeking to avoid errors from taking observed values as optimal, the approach may substitute unachievable output / average cost pairs to calculate a long-run cost function. Studies that use this method should ensure that the measure used for capital is appropriate and that results yield plausible optimal capital values for several functional forms and specifications.

Regarding scope economies, Chapter 5 shows that scope economies may exist where cost complementarities are absent. The possible presence of scope economies without cost complementarities is established as a theoretical point, but this thesis demonstrates a concrete example using data from England. The existence of scope economies without associated cost complementarities may be more likely where there are high fixed costs of production or complex interactions between multiple or heterogeneous outputs. As demonstrated in the high intercept term in the estimated cost functions, the production of hospital healthcare involves high fixed costs and is characterised by heterogeneous outputs. This thesis demonstrates limitations with approaches which rely on cost complementarities to investigate scope economies. The failure to detect scope economies by using such methods is not solely a theoretical concern but a demonstrable weakness.

6.4.3 Strengths and Weaknesses Summary

Strengths and weaknesses are discussed throughout the thesis. In summary, this thesis takes advantage of additional data made available in the last few years, making more detailed cost models. Many prior studies lacked access to input prices and relied on assumptions that national procurement and wage-setting allowed the omission of input prices. Using published wage rates and a constructed measure of capital, studies such as this one can start to account for differences in input prices faced by each hospital. Equally, the depth of DRG (HRG) coding allows output data from the UK to be weighted to a much more detailed level than prior studies. This study can therefore account for the heterogeneity in hospital output in a much more complex way than was previously the case.

Weaknesses include the lack of information on non-wage consumable inputs, such as medicines and other consumables used up in production. The lack of data on these inputs means it is difficult to model the effects of differential consumable prices on costs. Whilst some consumables and medicines are subject to national procurement, there will be local differentiation on other consumable input prices. Data on the cost and usage of agency staff is also lacking at the national level, which means that the effective wage rates for labour used in this study are for directly employed staff only. Hospitals that use significant amounts of indirectly employed agency staff may be poorly modelled until this data is available.

6.4.4 Future Research

Future research in this area is needed to better understand the variation in results caused by using different functional forms. In this thesis, the translog functional form gave scale economy estimates with less variation. Estimates also tended to decline with increased bed numbers. Quadratic functional forms tended to yield scale economy results with more variation and a weaker relation to size or output. Further research with other data sources could establish whether this is generally true and whether translog or quadratic functional forms ought to be preferred. Studies that aim to estimate limits to positive economies of scale could well be affected by the choice of functional form and may be advised to use both forms, at least as a sensitivity test.

Further work could also evaluate whether certain functional forms are more vulnerable to outlier observations, violations of convexity or other desirable economic properties. This work could also be combined with studies examining the effects of different specifications. More data on UK hospital costs, outputs and wages have become available in recent years. More detailed cost function specifications are possible but pose particular questions about the trade-off between model parsimony and accuracy. Further work could be conducted to investigate the most appropriate means of aggregating heterogeneous outputs and inputs to better model hospital costs.

This study also found that approaches which use an envelope approach to generate a long-run cost function may also be vulnerable to the poor quality of proxy measures for capital, leading to ‘optimal capital values’ which are implausible or possibly biased. In this thesis, the quadratic functional form was particularly vulnerable to this, perhaps due to a failure of convexity and the lack of a global minimum point. Future research could examine whether bed numbers remain a good proxy measure of economic capital and, if not, which measures could be employed. Given the growth in ambulatory care and the management of chronic conditions by specialists, more nuanced proxy measures for capital may need to be developed and tested if approaches based on an envelope methodology are to be used. Studies using envelope methods may need to demonstrate that bed numbers or other proxy measures of capital are appropriate.

Future research could also look in more detail at economies of scope. This thesis contains a short evaluation showing that General Surgery particularly exhibits scope economies, as well as General Medicine and Obstetrics / Gynaecology to a lesser extent. There are some *a priori* reasons to expect more 'general' specialties to exhibit scope economies. These specialties may be particularly involved in the joint production of healthcare with other specialties. Alternately, these specialties may have clinical labour more easily substituted for roles elsewhere. However, this study does not evaluate these hypotheses. Future research into why scope economies are observed in these specialties would be useful. This additional research could take the form of pairwise comparisons to explore which particular combinations of specialties have scope economies. Alternatively, it may be that the source of scope economies lies more in the general sharing of fixed costs across all types of activity, with some specialties taking advantage of this more than others.

6.5 Summary

This thesis offers novel evidence on hospital healthcare economies of scale, using data from the English NHS. It demonstrates that small but positive scale economies are present up to a higher limit (1,200 beds) than older and more dated studies. For larger hospitals, the limit at which scale economies are no longer positive varies according to the functional form and estimation method. The thesis also demonstrates differences in geographical regions driven by differences in wage rates. High-cost areas with high wage rates have lower scale economies as increases in output cause higher unit costs. Wage rates and input prices are often omitted from similar studies, using the rationale that hospitals face the same input prices. However, hospitals may not always face the same input prices, and studies that make similar assumptions may be unable to model differences in costs arising from input price variation.

Policymakers concerned with hospital costs and output can note that whilst the degree of scale economies observed is unlikely to justify extensive reorganisation, some gains could be realised from expansion for most English hospitals. There is little evidence that diseconomies of scale are present at any level of size. Policymakers can also note that certain specialties, notably General Surgery, demonstrate scope economies with other specialties and should therefore be present in each hospital to maximise positive spillovers. These particular specialties ought not to be centralised in larger hospital trusts.

Researchers working in this area can take the main estimation results of this thesis as another piece of evidence in the literature on hospital scale economies, which has become dated in recent years. This thesis also demonstrates the variation in results that can arise from differences in functional form and method and offers some examples of how results can be affected by differences in method. Future research could explore how differences in method contribute to differences in results and further explore scope economies, which remains an underresearched area.

This area of research is an important area with many practical policy implications. The heterogeneous nature of hospital healthcare outputs and inputs means that researchers modelling cost functions must make tradeoffs between parsimony and accurate estimation. However, it is also an area where data is becoming increasingly available, with the adoption of electronic records and other developments reducing the cost of collecting and producing data. More research in this area could provide important results for those interested in the efficient allocation of healthcare resources in hospitals.

Appendix A

Excluded Specialist Organisations

Organisation	Specialty
Liverpool Heart and Chest	Cardiothoracic
Papworth	Cardiothoracic
Royal Brompton	Cardiothoracic
The Walton	Neurology
The Christie	Oncology
The Clatterbridge	Oncology
Royal Marsden	Ophthalmology
Royal Moorefields	Ophthalmology
Robert Jones Orthopaedic	Orthopaedic
Royal National Orthopaedic Hospital	Orthopaedic
The Royal Orthopaedic hospital	Orthopaedic
Alder Hey Children's NHS Foundation Trust	Paediatric
Great Ormond Street	Paediatric
Sheffield Children's NHS Foundation Trust	Paediatric
Queen Victoria Hospital	Reconstructive Surgery & Rehabilitation
Royal National Hospital for Rheumatic Diseases	Rheumatic Diseases
Liverpool Women's	Women's
Birmingham Women's	Women's & Paediatric

Appendix B

Simplified Output Model (envelope method)

As a robustness check to the main result in Chapter 4, it is useful to test the sensitivity of the optimal beds calculation to different specifications and configurations of outputs. This check could be particularly important when we use beds as the proxy measure for capital. Required bed numbers in each hospital may be affected by the types of care provided, so calculated optimal values may be affected by the cost function specification and how outputs are defined. Studies such as Kristensen et al. (2012) often aggregate outputs into one category to deal with complications arising from the multi-input case. However, doing so may obscure qualitative differences in the types of healthcare provided. These qualitative differences may affect the observed number of beds. Comparing results under an aggregated specification is useful to see the effect of the chosen output specification on scale economy calculations. Other authors that have tested alternate specifications (Butler, 1995, ch 6) have found scale economy estimates were sensitive to specification.

Firstly, an aggregated output measure is defined by multiplying each weighted output grouping by the average cost for that output group in the dataset. For example, the average cost for an A&E attendance is £156, so all HRG-weighted A&E activity is multiplied by this value. T. This calculation is repeated for all other categories, summing the results to obtain an HRG-adjusted activity figure in each category. The result is divided by the mean cost of an average elective spell (£1,259). The resulting aggregate output can be considered as 'elective inpatient equivalents' for each hospital.

In mathematical terms, aggregate output figures are therefore calculated as:

$$y_{agg} = \left(y_{ae} \cdot \overline{C_{ae}} + y_{cc} \cdot \overline{C_{cc}} + y_{ds} \ \cdot \ \overline{C_{ds}} + y_{el} \ \cdot \ \overline{C_{el}} + y_{ne} \ \cdot \ \overline{C_{nel}} + y_{op} \ \cdot \ \overline{C_{op}} \right) / \, \overline{C_{el}} \tag{B.1}$$

Where y_i represents the amount of output i produced after the within-category weighting described in Section 2.3.2, and $\overline{C_i}$ represents the average cost of that output group, as defined in Table 2.3. Subscripts in the equation denote the type of output; ae - Accident and Emergency, cc - Critical Care, ds - Diagnostic Services, el - Elective Inpatient, nel - Non-Elective Inpatient, op - Outpatient.

Descriptive statistics for the aggregate output measure are set out in Table B.1, showing how the aggregate output measure compares to the previous output categories.

Table B.1: Aggregate Output Measure Compared to Original Output Categories

Output Type	**Mean**	**SD**	**Median**	**Q1**	**Q4**
Aggregate Output	**247,185**	**118,729**	**217,673**	**160,540**	**308,081**
A&E Attendance	127,483	60,463	112,013	85,640	152,096
Critical Care	19,988	19,585	12,709	8,016	21,496
Diagnostic Services	4,297,133	2,508,860	3,777,083	2,642,614	5,255,448
Elective Inpatient	51,604	25,862	46,910	31,734	64,884
Non-Elective Inpatient	71,745	34,026	65,122	47,133	90,191
Outpatient	591,701	293,106	530,505	377,434	738,400

N =138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19] hospitals

This composite value is used in a single-output quadratic specification along with input prices for training grade doctor and nursing wages, the cost of capital variable, and average bed numbers. Coefficients for the directly estimated regression are shown in Table B.2. In this regression, the exponent of the intercept value is the expected total cost for a non-teaching hospital (£248,520,469), where aggregate output, training grade doctor and nursing wages equal the mean for the dataset. The coefficient on the aggregate output term (0.85) reflects the marginal percentage change in total cost from an additional 1% increase in output (at the sample mean). Coefficients on input price variables likewise reflect the marginal percentage effect on total cost from a 1% increase in input wages. The teaching hospital coefficient (8%) shows the expected percentage increase in total cost due to a hospital engaging in teaching.

The positive coefficient on the squared output term suggests that costs rise more than proportionately as output increases past the sample mean. Other statistically significant terms included the squared term for nurse wages, suggesting that total costs become more responsive to nursing wages as they increase. A similarly significant relationship was observed for the capital cost term, whilst that for training grade doctors is negative but not statistically significant. The interaction between nursing wages and training grade doctors is statistically significant and negative, suggesting that as wages for one group, the relative impact on the cost of the other declines.

Table B.2: Estimation results - Directly Estimated Translog Form (Aggregated Output)

Term	**Estimate**	**Std. Error**
Intercept	19.331	(0.017)***
Aggregated Output	0.853	(0.017)***
Training Grade Doctor Wages	0.405	(0.076)***
Nurse Wages	0.302	(0.129)*
Cost of Capital	0.005	(0.009)
0.5 * Aggregated Output * Aggregated Output	0.089	(0.041)*
Aggregated Output * Training Grade Doctor Wages	-0.049	(0.159)
Aggregated Output * Nurse Wages	0.209	(0.227)
Aggregated Output * Cost of Capital	0.008	(0.016)
0.5 * Training Grade Doctor Wages * Training Grade Doctor Wages	-0.503	(1.537)
Nurse Wages * Training Grade Doctor Wages	-4.030	(1.861)*
0.5 * Nurse Wages * Nurse Wages	7.256	(3.063)*
Cost of Capital * Training Grade Doctor Wages	0.197	(0.122)
Cost of Capital * Nurse Wages	0.083	(0.140)
0.5 * Cost of Capital * Cost of Capital	0.045	(0.019)*
Teaching Hospital Organisation	0.077	(0.025)**

R^2 = 0.9758 adj R^2 = 0.9754 AIC = -2,667 BIC = -2,592 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

Including beds in the model to calculate a short-run cost function results in a short-run cost estimation shown in Table B.3. The coefficient on the beds variable is markedly different from the multi-output case. In this aggregated version, the coefficients on the beds term and the square of beds are 2.55 and -0.36, respectively.

The equivalent coefficients on the original multi-output model are 0.17 and 0.33. These differences imply substantially different optimal bed calculations even at low output levels, where the cross-product beds-output terms are less important. The sensitivity of the bed coefficients to changes in the specification of outputs suggests that optimal bed calculations may be similarly sensitive to the construction of the estimated cost function, particularly the definition and categorisation of outputs.

Table B.3: Estimation results - Short-run Translog Form (Aggregated Output)

Term	**Estimate**	**Std. Error**
Intercept	10.328	(2.942)***
Aggregated Output	-0.110	(0.837)
Training Grade Doctor Wages	-6.077	(2.328)**
Nurse Wages	9.004	(2.731)***
Cost of Capital	-0.081	(0.273)
0.5 * Aggregated Output * Aggregated Output	0.021	(0.132)
0.5 * Average Beds * Average Beds	-0.358	(0.136)**
Average Beds	2.550	(0.894)**
Average Beds * Aggregated Output	0.133	(0.127)
Average Beds * Training Grade Doctor Wages	0.980	(0.352)**
Average Beds * Nurse Wages	-1.306	(0.414)**
Average Beds * Cost of Capital	0.013	(0.041)
Aggregated Output * Training Grade Doctor Wages	-0.639	(0.311)*
Aggregated Output * Nurse Wages	1.067	(0.369)**
Aggregated Output * Cost of Capital	-0.001	(0.037)
0.5 * Training Grade Doctor Wages * Training Grade Doctor Wages	-1.788	(1.525)
Nurse Wages * Training Grade Doctor Wages	-4.309	(1.850)*
0.5 * Nurse Wages * Nurse Wages	7.642	(2.994)*
Cost of Capital * Training Grade Doctor Wages	0.219	(0.119)
Cost of Capital * Nurse Wages	0.035	(0.138)
0.5 * Cost of Capital * Cost of Capital	0.041	(0.019)*
Teaching Hospital Organisation	0.057	(0.022)**

R^2 = 0.9827 adj R^2 = 0.9823 AIC = -2,698 BIC = -2,595 N = 798 total observations from 141 hospitals, observations split as follows: 138 [2013/14], 134 [2014/15], 134 [2015/16], 133 [2016/17], 132 [2017/18], 127 [2018/19]

*** - Statistically Significant at 5%, ** - Statistically Significant at 1%, *** - Statistically Significant at 0.1%**

Optimal beds calculated under this alternate specification differ from the main specification computed in Figure 4.1. The range of the distribution is similar, but optimal beds calculated under this aggregated output specification are higher than the disaggregated main version. The partial derivative of the log of total cost with respect to the log of beds is:

$$\begin{aligned}\frac{\partial \ln\, C(y,w,K')}{\partial \ln\, K'} =& -0.358 \cdot \ln\, Beds + 2.55\ + 0.133 \cdot \ln\, Aggregate\, Output \\ &+ 1.390 \cdot \ln\, Training\, Grade\, Doctor\, Wages \\ &+ 0.003 \cdot \ln\, Nursing\, Wages + 0.470 \cdot \ln\, Cost\, of\, Capital\end{aligned} \tag{B.2}$$

Setting equal to zero and rearranging, this gives the log of optimal capital stock $\ln\, K'$ (in terms of beds) as:

$$\begin{aligned}\ln\, Beds =& 7.116 + 0.371 \cdot \ln\, Aggregate\, Output \\ &+ 2.735 \cdot \ln\, Training\, Grade\, Doctor\, Wages \\ &- 3.645 \cdot \ln\, Nursing\, Wages\ + 0.035 \cdot \ln\, Cost\, of\, Capital\end{aligned} \tag{B.3}$$

When optimal beds are calculated and plotted against actual beds (Figure B.1), the simplified or aggregated specification returns much higher optimal bed levels.

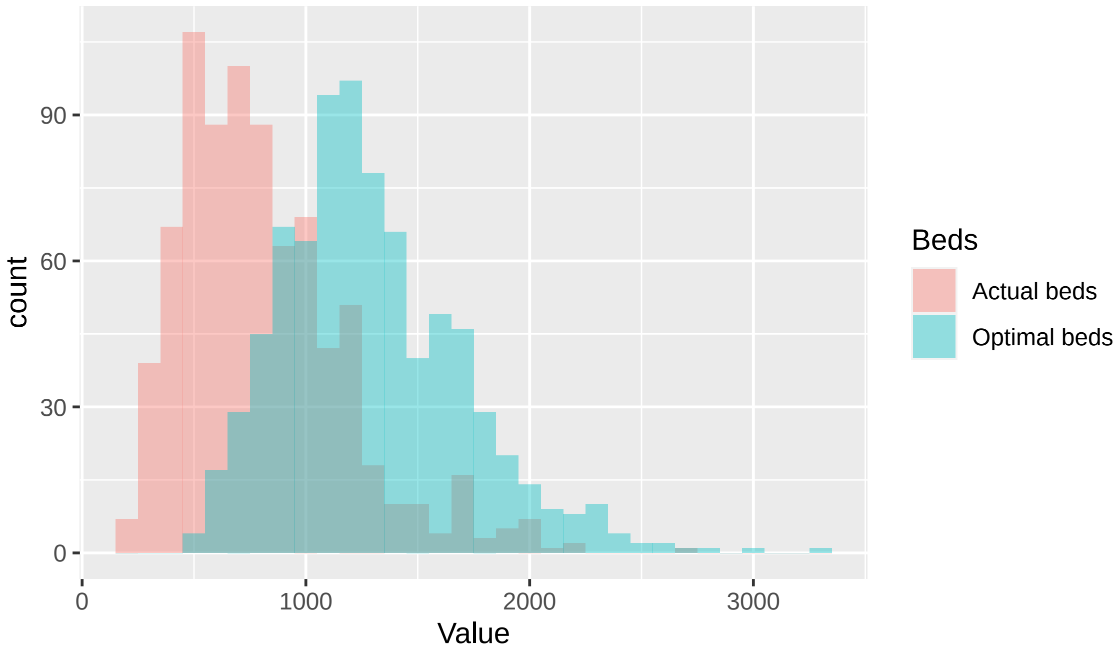

Figure B.1: Distribution of calculated optimal beds and observed beds, aggregated output specification. See Fig 4.1 for details on derivation and interpretation.

Scale economy indices for this simplified specification are similar overall to the main result, with constant economies of scale obtained with a direct estimation, more positive economies of scale observed in the short-run and lower scale economies measured in the long-run. Figure B.2 shows the scale economy calculations for a directly estimated simplified model (no beds included), a short-run cost version including beds, and a long-run version substituting calculated optimal beds into the cost function. As with the main translog specification, direct estimation tends to yield the lowest scale economy estimations with a median scale economy index of 1.18. Short-run scale economies returned the highest values (median value 1.32), and long-run estimation between the two (median value 1.31).

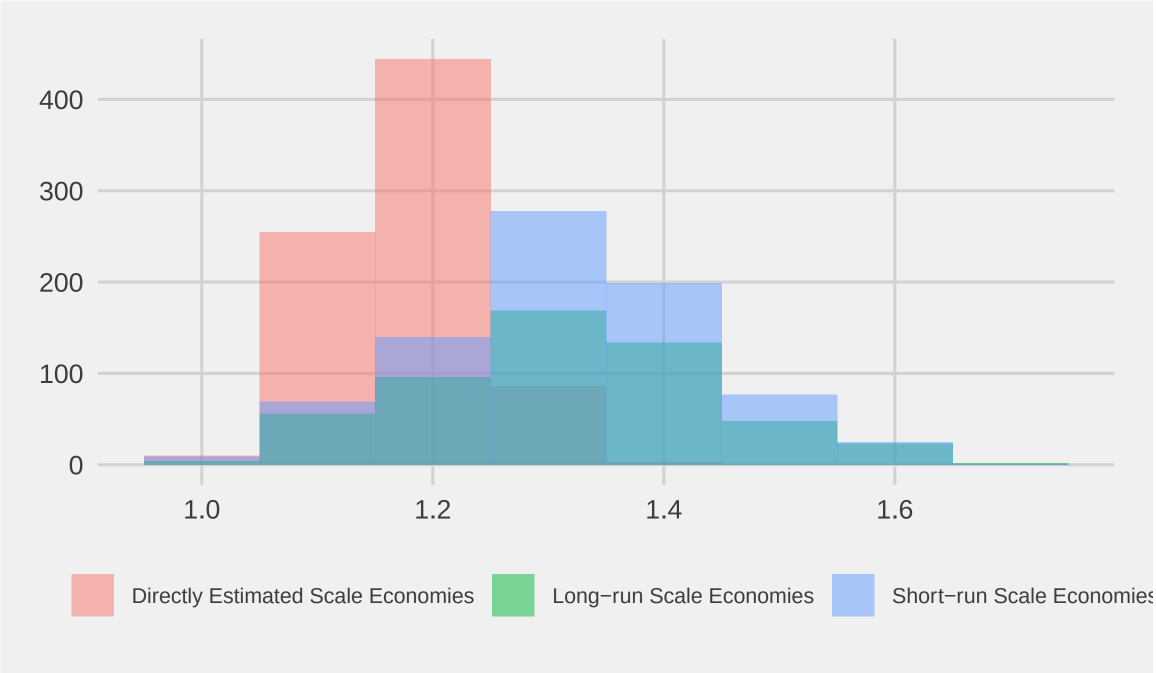

Figure B.2: Distribution of Scale Economies (Aggregated Output). Scale economies are estimated using the direct estimation method.

The relationships between each version of scale economies and size measured in beds (shown in Figures B.3 - B.6) are comparable to those obtained by the standard specification. Figure B.3 plots directly estimated scale economies against beds. Recorded scale economies and their relationship to changes in bed numbers are similar to the main result shown in Figure 2.19, though the simplified version has lower scale economies for the few observations with more than 1,000 beds.

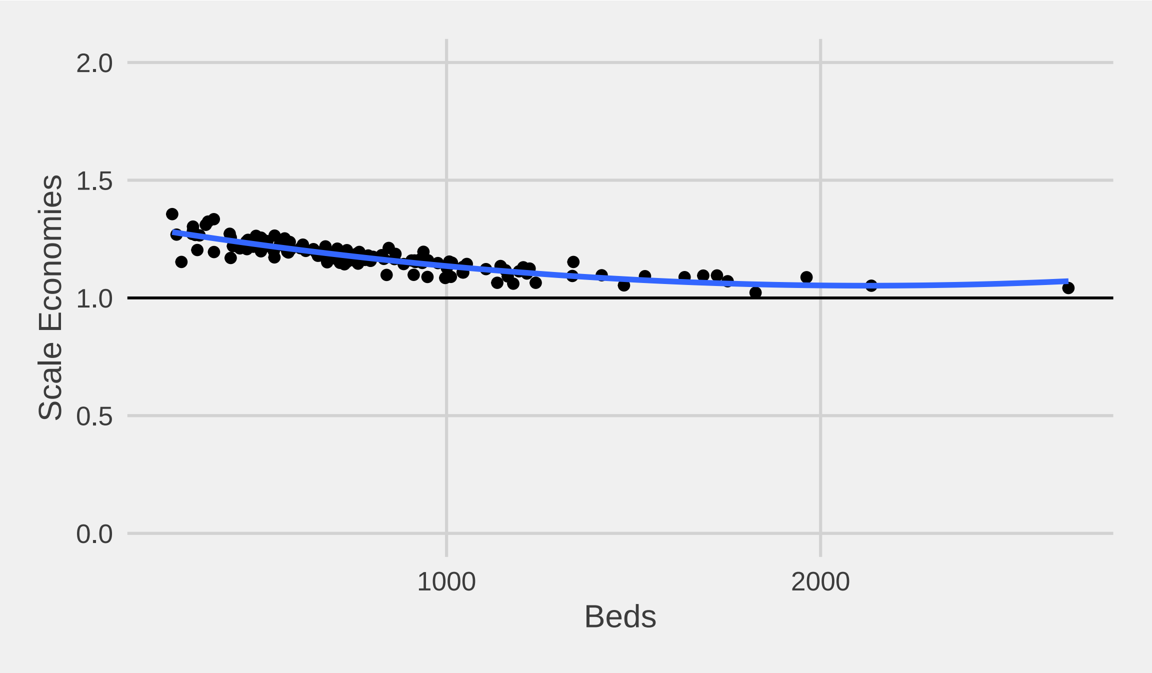

Figure B.3: Scale Economies vs Current Beds (Aggregated Output). Scale economies are estimated using the direct estimation method.

The relationship between short-run scale economies and bed numbers, shown in Figure B.4, is also similar to the main short-run result (Figure 4.3). Short-run scale economies are generally positive and higher than direct estimation, declining with size until reaching a rough limit at 1,250 beds, where most observations exhibit constant or negative scale economies.

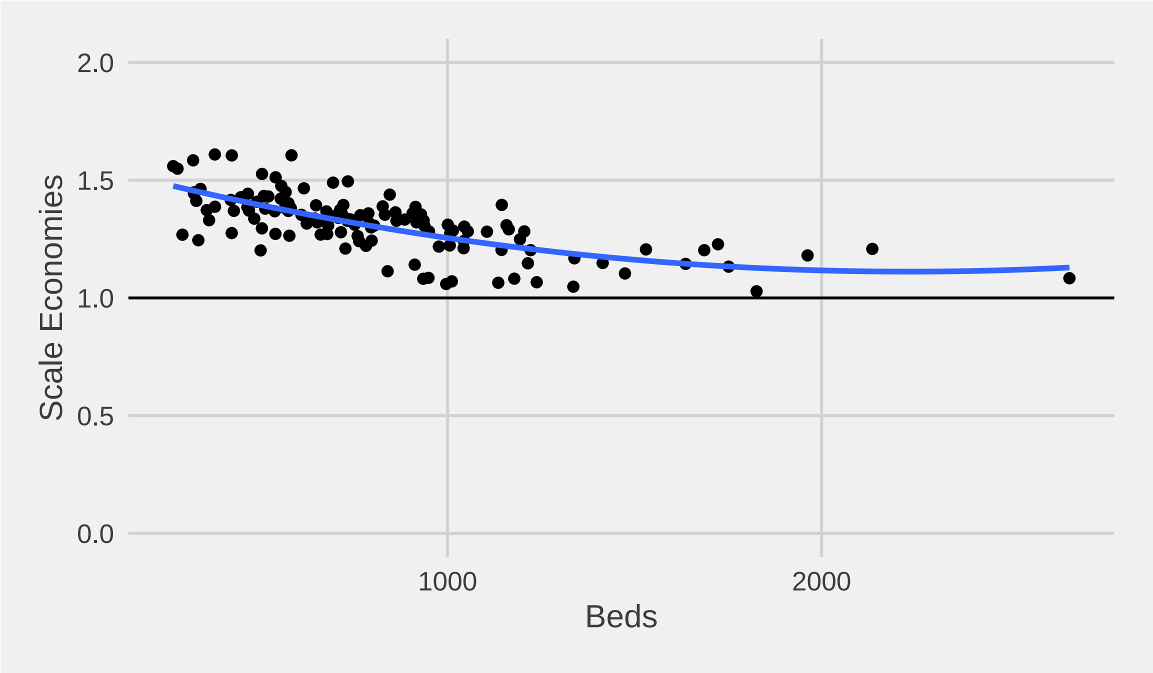

Figure B.4: Short-Run Scale Economies vs Current Beds (Aggregated Output).

Figures B.5 and B.6 show the long-run scale economy calculations, with each observation assumed to have optimal capital stock for its current output level. Again, the simplified translog specification shows similar results to the main estimation. Scale economies are highest amongst small trusts and decline with increasing size. This relationship is observed for both current and calculated optimal beds.

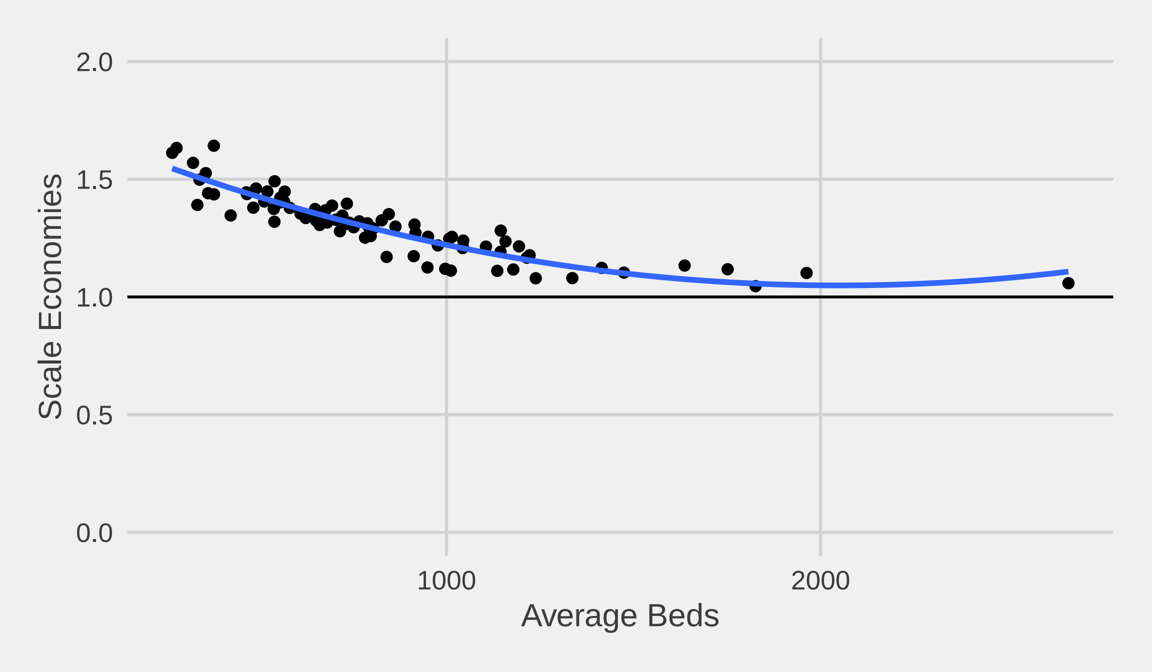

Figure B.5: Long-Run Scale Economies vs Current Beds (Aggregated Output).

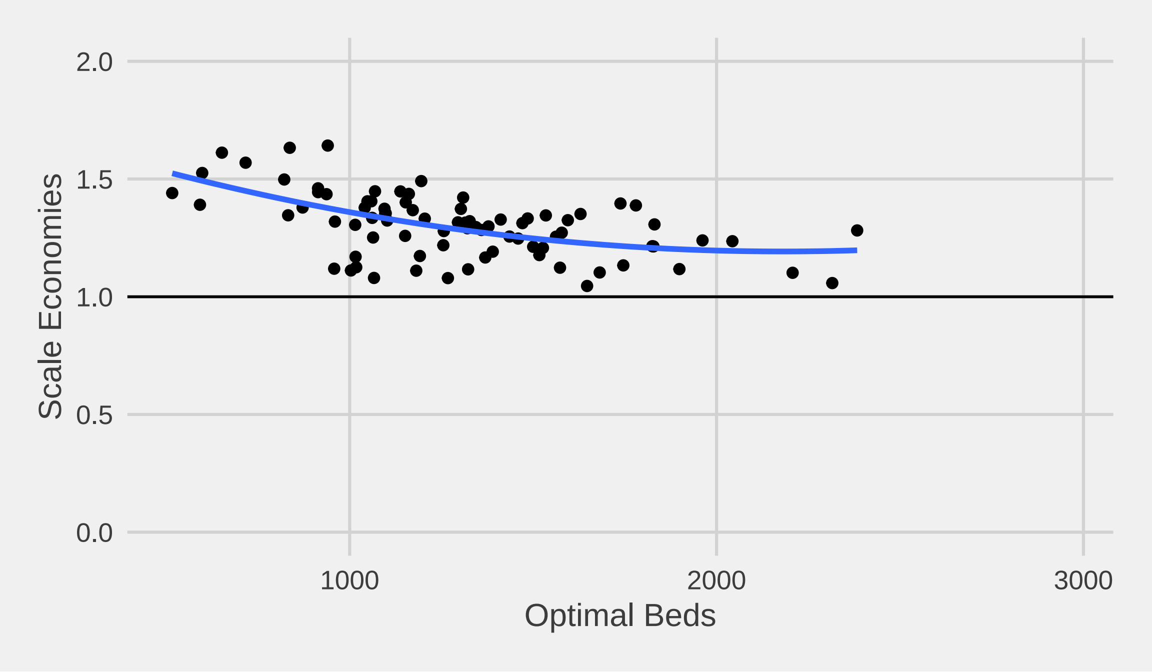

Figure B.6: Long-Run Scale Economies vs Optimal Beds (Aggregated Output)

Appendix C

Specialty Groupings

Service Code	Description	Specialty Grouping
100	General Surgery Service	General Surgery
101	Urology Service	Other
102	Transplant Surgery Service	General Surgery
103	Breast Surgery Service	General Surgery
104	Colorectal Surgery Service	General Surgery
105	Hepatobiliary And Pancreatic Surgery Service	General Surgery
106	Upper Gastrointestinal Surgery Service	General Surgery
107	Vascular Surgery Service	Other
108	Spinal Surgery Service	Orthopaedics
110	Trauma And Orthopaedic Service	Orthopaedics
111	Orthopaedic Service	Orthopaedics
120	Ear Nose And Throat Service	Other
130	Ophthalmology Service	Other
140	Oral Surgery Service	Other
141	Restorative Dentistry Service	Other
142	Paediatric Dentistry Service	Other
143	Orthodontic Service	Other
144	Maxillofacial Surgery Service	Other
145	Oral And Maxillofacial Surgery Service	Other
150	Neurosurgical Service	Other
160	Plastic Surgery Service	Other
161	Burns Care Service	Other
170	Cardiothoracic Surgery Service	Other
171	Paediatric Surgery Service	Other
172	Cardiac Surgery Service	Other
173	Thoracic Surgery Service	Other

Service Code	Description	Specialty Grouping
174	Cardiothoracic Transplantation Service	Other
180	Emergency Medicine Service	Emergency Medicine (including A&E)
190	Anaesthetic Service	Anaesthetics, Critical Care & Pain Management
191	Pain Management Service	Anaesthetics, Critical Care & Pain Management
192	Intensive Care Medicine Service	Anaesthetics, Critical Care & Pain Management
211	Paediatric Urology Service	Other
212	Paediatric Transplantation Surgery Service	General Surgery
213	Paediatric Gastrointestinal Surgery Service	General Surgery
214	Paediatric Trauma And Orthopaedic Service	Orthopaedics
215	Paediatric Ear Nose And Throat Service	Other
216	Paediatric Ophthalmology Service	Other
217	Paediatric Oral And Maxillofacial Surgery Service	Other
218	Paediatric Neurosurgery Service	Other
219	Paediatric Plastic Surgery Service	Other
220	Paediatric Burns Care Service	Other
221	Paediatric Cardiac Surgery Service	Other
222	Paediatric Thoracic Surgery Service	Other
223	Paediatric Epilepsy Service	Other
241	Paediatric Pain Management Service	Anaesthetics, Critical Care & Pain Management
242	Paediatric Intensive Care Service	Anaesthetics, Critical Care & Pain Management
250	Paediatric Hepatology Service	Other
251	Paediatric Gastroenterology Service	Other
252	Paediatric Endocrinology Service	Other
253	Paediatric Haematology Service	Other
254	Paediatric Audio Vestibular Medicine Service	Other
255	Paediatric Clinical Immunology And Allergy Service	Other
256	Paediatric Infectious Diseases Service	Other
257	Paediatric Dermatology Service	Other
258	Paediatric Respiratory Medicine Service	Other
259	Paediatric Nephrology Service	Other
260	Paediatric Medical Oncology Service	Other
261	Paediatric Inherited Metabolic Medicine Service	Other

Service Code	Description	Specialty Grouping
262	Paediatric Rheumatology Service	Other
263	Paediatric Diabetes Service	Other
280	Paediatric Interventional Radiology Service	Other
290	Community Paediatric Service	Other
291	Paediatric Neurodisability Service	Other
300	General Internal Medicine Service	General Medicine
301	Gastroenterology Service	Other
302	Endocrinology Service	Other
303	Haematology Service	Other
304	Clinical Physiology Service	Other
305	Clinical Pharmacology Service	Other
306	Hepatology Service	Other
307	Diabetes Service	Other
308	Blood And Marrow Transplantation Service	Other
309	Haemophilia Service	Other
310	Audio Vestibular Medicine Service	Other
311	Clinical Genetics Service	Other
313	Clinical Immunology And Allergy Service	Other
314	Rehabilitation Medicine Service	Other
315	Palliative Medicine Service	Other
316	Clinical Immunology Service	Other
317	Allergy Service	Other
318	Intermediate Care Service	Other
319	Respite Care Service	Other
320	Cardiology Service	Other
321	Paediatric Cardiology Service	Other
322	Clinical Microbiology Service	Other
323	Spinal Injuries Service	Other
324	Anticoagulant Service	Other
325	Sport And Exercise Medicine Service	Other
327	Cardiac Rehabilitation Service	Other
328	Stroke Medicine Service	Other
329	Transient Ischaemic Attack Service	Other
330	Dermatology Service	Other
331	Congenital Heart Disease Service	Other
333	Rare Disease Service	Other
340	Respiratory Medicine Service	Other
341	Respiratory Physiology Service	Other

Service Code	Description	Specialty Grouping
342	Pulmonary Rehabilitation Service	Other
344	Complex Specialised Rehabilitation Service	Other
345	Specialist Rehabilitation Service	Other
346	Local Specialist Rehabilitation Service	Other
350	Infectious Diseases Service	Other
352	Tropical Medicine Service	Other
360	Genitourinary Medicine Service	Other
361	Renal Medicine Service	Other
370	Medical Oncology Service	Other
371	Nuclear Medicine Service	Other
400	Neurology Service	Other
401	Clinical Neurophysiology Service	Other
410	Rheumatology Service	Other
420	Paediatric Service	Other
421	Paediatric Neurology Service	Other
422	Neonatal Critical Care Service	Anaesthetics, Critical Care & Pain Management
430	Elderly Medicine Service	Other
450	Dental Medicine Service	Other
460	Medical Ophthalmology Service	Other
501	Obstetrics Service	Obstetrics & Gynaecology
502	Gynaecology Service	Obstetrics & Gynaecology
503	Gynaecological Oncology Service	Obstetrics & Gynaecology
560	Midwifery Service	Obstetrics & Gynaecology
650	Physiotherapy Service	Other
651	Occupational Therapy Service	Other
652	Speech And Language Therapy Service	Other
653	Podiatry Service	Other
654	Dietetics Service	Other
655	Orthoptics Service	Other
656	Clinical Psychology Service	Other
657	Prosthetics Service	Other
658	Orthotics Service	Other
659	Dramatherapy Service	Other
662	Optometry Service	Other
663	Podiatric Surgery Service	Other
710	Adult Mental Health Service	Other
711	Child And Adolescent Psychiatry Service	Other

Service Code	Description	Specialty Grouping
712	Forensic Psychiatry Service	Other
713	Medical Psychotherapy Service	Other
715	Old Age Psychiatry Service	Other
720	Eating Disorders Service	Other
721	Addiction Service	Other
722	Liaison Psychiatry Service	Other
723	Psychiatric Intensive Care Service	Other
724	Perinatal Mental Health Service	Other
727	Dementia Assessment Service	Other
800	Clinical Oncology Service	Other
811	Interventional Radiology Service	Other
812	Diagnostic Imaging Service	Diagnostic & Pathology Services
822	Chemical Pathology Service	Diagnostic & Pathology Services
834	Medical Virology Service	Other
840	Audiology Service	Other
920	Diabetic Education Service	Other
999	Unknown	Other
CCU01	Non-Specific, General Adult Critical Care Patients Predominate	Anaesthetics, Critical Care & Pain Management
CCU02	Surgical Adult Patients (Unspecified Specialty)	Anaesthetics, Critical Care & Pain Management
CCU03	Medical Adult Patients (Unspecified Specialty)	Anaesthetics, Critical Care & Pain Management
CCU04	Paediatric Intensive Care Unit (Paediatric Critical Care Patients Predominate)	Anaesthetics, Critical Care & Pain Management
CCU05	Neurosciences Adult Patients Predominate	Anaesthetics, Critical Care & Pain Management
CCU06	Cardiac Surgical Adult Patients Predominate	Anaesthetics, Critical Care & Pain Management
CCU07	Thoracic Surgical Adult Patients Predominate	Anaesthetics, Critical Care & Pain Management
CCU08	Burns And Plastic Surgery Adult Patients Predominate	Anaesthetics, Critical Care & Pain Management
CCU09	Spinal Adult Patients Predominate	Anaesthetics, Critical Care & Pain Management
CCU10	Renal Adult Patients Predominate	Anaesthetics, Critical Care & Pain Management

Service Code	Description	Specialty Grouping
CCU11	Liver Adult Patients Predominate	Anaesthetics, Critical Care & Pain Management
CCU12	Obstetric And Gynaecology Critical Care Patients Predominate	Anaesthetics, Critical Care & Pain Management
CCU13	Neonatal Intensive Care Unit	Anaesthetics, Critical Care & Pain Management
CCU14	Facility For Babies On A Transitional Care Ward	Anaesthetics, Critical Care & Pain Management
CCU15	Facility For Babies On A Maternity Ward	Anaesthetics, Critical Care & Pain Management
CCU16	Ward For Children And Young People	Anaesthetics, Critical Care & Pain Management
CCU17	High Dependency Unit For Children And Young People	Anaesthetics, Critical Care & Pain Management
CCU18	Renal Unit For Children And Young People	Anaesthetics, Critical Care & Pain Management
CCU19	Burns Unit For Children And Young People	Anaesthetics, Critical Care & Pain Management
CCU90	Non-Standard Location Using A Ward Area	Anaesthetics, Critical Care & Pain Management
CCU91	Non-Standard Location Using The Operating Department	Anaesthetics, Critical Care & Pain Management
CCU92	Non-Standard Location Using The Operating Department For Children And Young People.	Anaesthetics, Critical Care & Pain Management
DADS	Direct Access Diagnostic Services	Diagnostic & Pathology Services
DAPS	Direct Access Pathology Services	Diagnostic & Pathology Services
FPC	Family Planning Clinic	Other
HIV1	Hiv Or Aids, Category 1, New Patients	Other
HIV2	Hiv Or Aids, Category 2, Stable Patients	Other
HIV3	Hiv Or Aids, Category 3, Complex Patients	Other
NEO	Neonatal Critical Care	Anaesthetics, Critical Care & Pain Management
PD	Paediatric Critical Care	Anaesthetics, Critical Care & Pain Management
T01A	Type 01 Admitted	Emergency Medicine (including A&E)
T01NA	Type 01 Non Admitted	Emergency Medicine (including A&E)

Service Code	Description	Specialty Grouping
T02A	Type 02 Admitted	Emergency Medicine (including A&E)
T02NA	Type 02 Non Admitted	Emergency Medicine (including A&E)
T03A	Type 03 Admitted	Emergency Medicine (including A&E)
T03NA	Type 03 Non Admitted	Emergency Medicine (including A&E)
T04A	Type 04 Admitted	Emergency Medicine (including A&E)
T04NA	Type 04 Non Admitted	Emergency Medicine (including A&E)

Bibliography

10 Aletras, V. H. (1999). A comparison of hospital scale effects in short-run and long-run cost functions. *Health Economics*, *8*(6), 521–530. https://doi.org/10.1002/(sici)1099-1050(199909)8:6%3C521::aid-hec462%3E3.0.co;2-g

Asmild, M., Hollingsworth, B., & Birch, S. (2013). The scale of hospital production in different settings: One size does not fit all. *Journal of Productivity Analysis*, *40*(2), 197–206. https://doi.org/10.1007/s11123-012-0332-9

Azevedo, H., & Mateus, C. (2014). Cost effects of hospital mergers in Portugal. *The European Journal of Health Economics*, *15*(9), 999–1010. https://doi.org/10.1007/s10198-013-0552-6

Baumol, W. J., Panzar, J. C., & Willig, R. D. (1982). *Contestable markets and the theory of industry structure*. Harcourt Brace Jovanovich.

Begg, C. B., Cramer, L. D., Hoskins, W. J., & Brennan, M. F. (1998). Impact of hospital volume on operative mortality for major cancer surgery. *JAMA*, *280*(20), 1747–1751. https://doi.org/10.1001/jama.280.20.1747

Bell, C. R. (1988). Economies of, versus returns to, scale: A clarification. *The Journal of Economic Education*, *19*(4), 331–335. https://doi.org/10.2307/1182344

Berry, R. E. (1967). Returns to scale in the production of hospital services. *Health Serv Res*, *2*(2), 123–139.

Breyer, F. (1987). The specification of a hospital cost function: A comment on the recent literature. *Journal of Health Economics*, *6*(2), 147–157. https://doi.org/10.1016/0167-6296(87)90004-X

Brown, R. S., Caves, D. W., & Christensen, L. R. (1979). Modelling the Structure of Cost and Production for Multiproduct Firms. *Southern Economic Journal*, *46*(1), 256. https://doi.org/10.2307/1057018

Butler, J. R. (1995). *Hospital cost analysis* (Vol. 3). Springer Science & Business

Media.

Carey, K., Burgess, J. F., & Young, G. J. (2015). Economies of Scale and Scope: The Case of Specialty Hospitals. *Contemporary Economic Policy*, *33*(1), 104–117. https://doi.org/10.1111/coep.12062

Carey, K., & Stefos, T. (2011). Controlling for quality in the hospital cost function. *Health Care Management Science*, *14*(2), 125–134. https://doi.org/10.1007/s10729-010-9142-7

Caves, D. W., Christensen, L. R., & Tretheway, M. W. (1980). Flexible Cost Functions for Multiproduct Firms. *The Review of Economics and Statistics*, *62*(3), 477. https://doi.org/10.2307/1927120

Chambers, R. (1988). *Applied production analysis*. Cambridge University Press.

Clark, J. R. (2012). Comorbidity and the Limitations of Volume and Focus as Organizing Principles. *Medical Care Research and Review*, *69*(1), 83–102. https://doi.org/10.1177/1077558711418520

Conrad, R. F., & Strauss, R. P. (1983). A multiple-output multiple-input model of the hospital industry in north carolina. *Applied Economics*, *15*(3), 341. https://doi.org/10.1080/00036848300000005

Cowing, T. G., & Holtmann, A. G. (1983). Multiproduct Short-Run Hospital Cost Functions: Empirical Evidence and Policy Implications from Cross-Section Data. *Southern Economic Journal*, *49*(3), 637–653. https://doi.org/10.2307/1058706

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, *27*(2), 1–43. https://doi.org/10.18637/jss.v027.i02

Culyer, A. J., Wiseman, J., Drummond, M. F., & West, P. A. (1978). What accounts for the higher costs of teaching hospitals? *Social Policy & Administration*, *12*(1), 20–30. https://doi.org/10.1111/j.1467-9515.1978.tb00121.x

Diewert, W. E. (1971). An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function. *Journal of Political Economy*, *79*(3), 481–507. https://doi.org/10.1086/259764

Dranove, D. (1998). Economies of scale in non-revenue producing cost centers: Implications for hospital mergers. *Journal of Health Economics*, *17*(1), 69–83. https://doi.org/10.1016/S0167-6296(97)00013-1

Euroqol. (2022). *EQ-5D*. https://euroqol.org/

Evans, R. G. (1971). "Behavioural" Cost Functions for Hospitals. *The Canadian Journal of Economics / Revue Canadienne d'Economique*, *4*(2), 198–215. https://doi.org/10.2307/133526

Farsi, M., & Filippini, M. (2008). Effects of ownership, subsidization and teaching activities on hospital costs in Switzerland. *Health Economics*, *17*(3), 335–350. https://doi.org/10.1002/hec.1268

Fournier, G. M., & Mitchell, J. M. (1992). Hospital Costs and Competition for Services: A Multiproduct Analysis. *The Review of Economics and Statistics*, *74*(4), 627–634. https://doi.org/10.2307/2109376

Fragkiadakis, G., Doumpos, M., Zopounidis, C., & Germain, C. (2016). Operational and economic efficiency analysis of public hospitals in Greece. *Annals of Operations Research*, *247*(2), 787–806. https://doi.org/10.1007/s10479-014-1710-7

Frech, H. E., & Mobley, L. R. (1995). Resolving the impasse on hospital scale economics: A new approach. *Applied Economics*, *27*(3), 286–296. https://doi.org/10.1080/00036849500000112

Freeman, M., Savva, N., & Scholtes, S. (2020). Economies of scale and scope in hospitals: An empirical study of volume spillovers. *Management Science*, *67*(2), 673–697. https://doi.org/10.1287/mnsc.2019.3572

Friedman, M. (1955). Discussion Section in "Survey of the Empirical Evidence on Economies of Scale". In *Business concentration and price policy* (pp. 213–238). National Bureau of Economic Research, Inc.

Fuss, M. A., & Waverman, L. (1981). Regulation and the multiproduct firm: The case of telecommunications in Canada. In *Studies in public regulation* (pp. 277–328). The MIT Press.

Fuss, M., McFadden, D., & Mundlak, Y. (1978). A Survey of Functional Forms in the Economic Analysis of Production. In *History of Economic Thought Chapters* (Vol. 1). McMaster University Archive for the History of Economic Thought.

Gaughan, J., Gutacker, N., Grašič, K., Kreif, N., Siciliani, L., & Street, A. (2019). Paying for efficiency: Incentivising same-day discharges in the English NHS. *Journal of Health Economics*, *68*, 102226. https://doi.org/10.1016/j.jhealeco.2019.102226

Gaynor, M., Laudicella, M., & Propper, C. (2012). Can governments do it better?

Merger mania and hospital outcomes in the English NHS. *Journal of Health Economics*, *31*(3), 528–543. https://doi.org/10.1016/j.jhealeco.2012.03.006

Giancotti, M., Guglielmo, A., & Mauro, M. (2017). Efficiency and optimal size of hospitals: Results of a systematic search. *PLOS ONE*, *12*(3), e0174533. https://doi.org/10.1371/journal.pone.0174533

Giannakas, K., Tran, K. C., & Tzouvelekas, V. (2003). On the choice of functional form in stochastic frontier modeling. *Empirical Economics*, *28*(1), 75–100. https://doi.org/10.1007/s001810100120

Gorman, I. E. (1985). Conditions for economies of scope in the presence of fixed costs. *The RAND Journal of Economics*, *16*(3), 431–436. https://doi.org/10.2307/2555569

Goudie, R., & Goddard, M. (2011). *Review of evidence on what drives economies of scope and scale in the provision of NHS services, focusing on A&E and associated hospital services: A report for the OHE commission on competition in the NHS.* Centre for Health Economics, University of York.

Grannemann, T. W., Brown, R. S., & Pauly, M. V. (1986). Estimating hospital costs: A multiple-output analysis. *Journal of Health Economics*, *5*(2), 107–127. https://doi.org/10.1016/0167-6296(86)90001-9

Gravelle, H., & Rees, R. (2004). *Microeconomics.* Pearson education.

Gutacker, N., Bojke, C., Daidone, S., Devlin, N. J., Parkin, D., & Street, A. (2013). Truly Inefficient or Providing Better Quality of Care? Analysing the Relationship Between Risk-Adjusted Hospital Costs and Patients' Health Outcomes. *Health Economics*, *22*(8), 931–947. https://doi.org/10.1002/hec.2871

Gutacker, N., Siciliani, L., Moscelli, G., & Gravelle, H. (2016). Choice of hospital: Which type of quality matters? *Journal of Health Economics*, *50*, 230–246. https://doi.org/10.1016/j.jhealeco.2016.08.001

Ham, C. (2018). Making sense of integrated care systems, integrated care partnerships and accountable care organisations in the NHS in England. In *London: The King's Fund.*

Heathfield, D. F. (1987). *An introduction to cost and production functions.* Macmillan International Higher Education.

Hefty, T. R. (1969). Returns to scale in hospitals: A critical review of recent research. *Health Serv Res*, *4*(4), 267–280.

Hollingsworth, B. (2003). Non-Parametric and Parametric Applications Measuring Efficiency in Health Care. *Health Care Management Science*, *6*(4), 203–218. https://doi.org/10.1023/A:1026255523228

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics*, *17*(10), 1107–1128. https://doi.org/10.1002/hec.1391

Holmås, T. H., Kamrul Islam, M., & Kjerstad, E. (2013). Between two beds: Inappropriately delayed discharges from hospitals. *International Journal of Health Care Finance and Economics*, *13*(3-4), 201–217. https://doi.org/10.1007/s10754-013-9135-4

Iacobucci, D., Schneider, M. J., Popovich, D. L., & Bakamitsos, G. A. (2016). Mean centering helps alleviate “micro” but not “macro” multicollinearity. *Behavior Research Methods*, *48*(4), 1308–1317.

Keeler, T. E., & Ying, J. S. (1996). Hospital Costs and Excess Bed Capacity: A Statistical Analysis. *The Review of Economics and Statistics*, *78*(3), 470. https://doi.org/10.2307/2109794

Koenker, R. (1977). Optimal Scale and the Size Distribution of American Trucking Firms. *Journal of Transport Economics and Policy*, *11*(1), 54–67.

Kohl, S., Schoenfelder, J., Fügener, A., & Brunner, J. O. (2019). The use of Data Envelopment Analysis (DEA) in healthcare with a focus on hospitals. *Health Care Management Science*, *22*(2), 245–286. https://doi.org/10.1007/s10729-018-9436-8

Kristensen, T., Bogetoft, P., & Pedersen, K. M. (2010). Potential gains from hospital mergers in denmark. *Health Care Management Science*, *13*(4), 334–345. https://doi.org/https://link.springer.com/journal/volumesAndIssues/10729

Kristensen, T., Olsen, K. R., Kilsmark, J., Lauridsen, J. T., & Pedersen, K. M. (2012). Economies of scale and scope in the Danish hospital sector prior to radical restructuring plans. *Health Policy*, *106*(2), 120–126. https://doi.org/10.1016/j.healthpol.2012.04.001

Lave, J. R., & Lave, L. B. (1970). Hospital cost functions. *The American Economic Review*, *60*(3), 379–395. http://www.jstor.org/stable/1817988

Leibenstein, H. (1966). Allocative Efficiency vs. "X-Efficiency". *The American Economic Review*, *56*(3), 392–415.

Li, T., & Rosenman, R. (2001). Estimating hospital costs with a generalized Leontief function. *Health Economics*, *10*(6), 523–538. https://doi.org/10.1002/hec.605

Linna, M., & Häkkinen, U. (2006). Reimbursing for the costs of teaching and research in finnish hospitals: A stochastic frontier analysis. *International Journal of Health Care Finance and Economics*, *6*(1), 83–97. https://doi.org/10.1007/s10754-006-6256-z

Mann, J. K., & Yett, D. E. (1968). The Analysis of Hospital Costs: A Review Article. *The Journal of Business*, *41*(2), 191–202.

Marder, W. D., & Hough, D. E. (1983). Medical Residency as Investment in Human Capital. *The Journal of Human Resources*, *18*(1), 49–64. https://doi.org/10.2307/145656

McKee, M. (2004). *Reducing hospital beds: What are the lessons to be learned?* World Health Organization. Regional Office for Europe.

Moscelli, G., Siciliani, L., Gutacker, N., & Gravelle, H. (2016). Location, quality and choice of hospital: Evidence from England 2002–2013. *Regional Science and Urban Economics*, *60*, 112–124. https://doi.org/10.1016/j.regsciurbeco.2016.07.001

NHS Centre for Reviews and Dissemination. (1996). *Hospital volume and health care outcomes, costs and patient access.* (Vol. 2(8).). University of York.

NHS Digital. (2022a). *The why, what and how of Casemix.* https://digital.nhs.uk/services/national-casemix-office/the-why-what-and-how-of-casemix

NHS Digital. (2022b). *TREATMENT FUNCTION CODE.* https://www.datadictionary.nhs.uk/attributes/treatment_function_code.html?hl=treatment%2Cfunction%2Ccode

NHS England. (2019). *Statistical work areas: Bed availability and occupancy.* https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/

NHS England. (2022a). Hospital bed. In *NHS Data Dictionary.* https://www.datadictionary.nhs.uk/supporting_information/hospital_bed.html

NHS England. (2022b). *NHS England » National Cost Collection for the NHS.* https://www.england.nhs.uk/costing-in-the-nhs/national-cost-collection/

NHS England. (2022c). *NHS providers: Trust accounts consolidation (TAC) data publications.* https://www.england.nhs.uk/financial-accounting-and-reporting/nhs-providers-tac-data-publications/

NHS Improvement. (2019). *Reference Costs.* https://www.england.nhs.uk/costing-in-the-nhs/national-cost-collection/

O'Neill, L., Rauner, M., Heidenberger, K., & Kraus, M. (2008). A cross-national comparison and taxonomy of DEA-based hospital efficiency studies. *Socio-Economic Planning Sciences*, *42*(3), 158–189. https://doi.org/10.1016/j.seps.2007.03.001

Pett, Tristan, Cooper, James, & Lewis, James. (2020). *Healthcare expenditure, UK Health Accounts: 2018.* Office for National Statistics.

Preyra, C., & Pink, G. (2006). Scale and scope efficiencies through hospital consolidations. *Journal of Health Economics*, *25*(6), 1049–1068. https://doi.org/10.1016/j.jhealeco.2005.12.006

Prior, D. (1996). Technical efficiency and scope economies in hospitals. *Applied Economics*, *28*(10), 1295–1301. https://doi.org/10.1080/000368496327840

R Core Team. (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rogers, R., Williams, S., Jarman, B., & Aylin, P. (2005). "HRG drift" and payment by results. *BMJ*, *330*(7491), 563. https://doi.org/10.1136/bmj.330.7491.563

Royal College of Surgeons. (2022). *General surgery.* https://www.rcseng.ac.uk/news-and-events/media-centre/media-background-briefings-and-statistics/general-surgery/

Scott, A., & Parkin, D. (1995). Investigating hospital efficiency in the new NHS: The role of the translog cost function. *Health Economics*, *4*(6), 467–478. https://doi.org/10.1002/hec.4730040604

Siciliani, L. (2006). Estimating Technical Efficiency in the Hospital Sector with Panel Data: A Comparison of Parametric and Non-parametric Techniques. *Applied Health Economics and Health Policy*, *5*(2), 99–116. https://link.springer.com/journal/volumesAndIssues/40258

Siciliani, L., Sivey, P., & Street, A. (2013). Differences in length of stay for hip replacement between public hospitals, specialised treatment centres and private providers: Selection or efficiency? *Health Economics*, *22*(2), 234–242. https:

//doi.org/10.1002/hec.1826

Siciliani, L., Stanciole, A., & Jacobs, R. (2009). Do waiting times reduce hospital costs? *Journal of Health Economics*, *28*(4), 771–780. https://doi.org/10.1016/j.jhealeco.2009.04.002

Sielska, A., & Nojszewska, E. (2022). Production function for modeling hospital activities. The case of Polish county hospitals. *PLOS ONE*, *17*(5), e0268350. https://doi.org/10.1371/journal.pone.0268350

Simar, L., & Wilson, P. W. (1998). Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models. *Management Science*, *44*(1), 49–61. https://doi.org/10.1287/mnsc.44.1.49

Sinay, U. A., & Campbell, C. R. (1995). Scope and scale economies in merging hospitals prior to merger. *Journal of Economics and Finance*, *19*(2), 107–123. https://doi.org/10.1007/BF02920513

Sloan, Frank A, Feldman, Roger D, & Steinwald, Bruce A. (1983). Effects of teaching on hospital costs. *Journal of Health Economics*, *2*(1), 1–28. https://doi.org/10.1016/0167-6296(83)90009-7

Smith, H., Currie, C., Chaiwuttisak, P., & Kyprianou, A. (2018). Patient choice modelling: How do patients choose their hospitals? *Health Care Management Science*, *21*, 259–268. https://doi.org/10.1007/s10729-017-9399-1

Tsai, P. F., & Mar Molinero, C. (2002). A variable returns to scale data envelopment analysis model for the joint determination of efficiencies with an example of the UK health service. *European Journal of Operational Research*, *141*(1), 21–38. https://doi.org/10.1016/S0377-2217(01)00223-5

Vita, M. G. (1990). Exploring Hospital Production Relationships with Flexible Functional Forms. *Journal of Health Economics*, *9*(1), 1–21. https://doi.org/http://www.sciencedirect.com/science/journal/01676296

Vitaliano, D. F. (1987). On the estimation of hospital cost functions. *Journal of Health Economics*, *6*(4), 305–318. https://doi.org/10.1016/0167-6296(87)90018-X

Wagstaff, A., & López, G. (1996). Hospital costs in Catalonia: A stochastic frontier analysis. *Applied Economics Letters*, *3*(7), 471–474. https://doi.org/10.1080/758540809

Weaver, M., & Deolalikar, A. (2004). Economies of scale and scope in Vietnamese hospitals. *Social Science & Medicine*, *59*(1), 199–208. https://doi.org/10.

1016/j.socscimed.2003.10.014

WHO. (2022). *ICD-10 version:2019*. https://icd.who.int/browse10/2019/en

Wilson, P. W., & Carey, K. (2004). Nonparametric Analysis of Returns to Scale in the US Hospital Industry. *Journal of Applied Econometrics*, *19*(4), 505–524. https://doi.org/10.1002/jae.801

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.