1 *Efficiency in health care*

1.1 Introduction

THE pursuit of efficiency has become a central objective of policy makers within most health systems. The reasons are manifest. In developed countries, expenditure on health care amounts to a sizeable proportion of gross domestic product. Policy makers need to be assured that such expenditure is in line with citizens' preferences, particularly when many sources of finance, such as tax revenues, are under acute pressure. On the supply side, health technologies are changing rapidly, and the pressures to introduce new technologies are often irresistible, even when there is uncertainty about cost-effectiveness. On the demand side, aging populations pose challenges for the design of health systems, and expectations are becoming ever more challenging. Finally, the revolution in information systems has made it feasible to measure aspects of system behaviour – most notably clinical activity – that until recently defied meaningful quantification.

The international concern was crystallised in the *World Health Report 2000* produced by the World Health Organization, which was devoted to the determinants and measurement of health system efficiency (World Health Organization 2000). The report stimulated a wide-ranging international debate, and a great deal of controversy (Williams 2001; Anand *et al.* 2002). However, its enduring legacy may be that it has helped policy makers to focus on the objectives of their health systems, on how achievement might be measured, and on whether resources are being deployed efficiently. A subsequent international conference organised by the Organization for Economic Co-operation and Development has confirmed the universal policy concern with performance measurement issues in health care (Smith 2002).

The analysis and measurement of efficiency is a complex undertaking, especially when there exist conceptual challenges, multiple objectives and great scope for measurement error. To address this complexity there has developed a flourishing research discipline of organisational efficiency analysis. Following pioneering studies by Farrell (1957), statisticians, econometricians and management scientists have developed tools to a high level of analytic sophistication that seek to measure the productive efficiency of organisations and systems. This book examines some of the most important techniques currently available to measure the efficiency of systems and organisations. It seeks to offer a critical assessment of the strengths and limitations of such tools applied to health and health care.

Throughout much of the book we take the view that health care objectives are known and agreed, and much of the discussion also assumes that the relative value placed on each objective is known. In practice, objectives and priorities are highly contested, and often not stated explicitly. A central purpose of this book is to examine how efficiency might be measured in the knowledge of objectives, but we also discuss the implications for efficiency analysis of failing to address priority setting explicitly.

Notwithstanding the apparent simplicity of the concept, there is a great deal of confusion in both popular and professional discussion about what is meant by efficiency in health care. In this opening chapter we first discuss the reasons for wishing to measure efficiency, and then define the concepts of organisational efficiency deployed in this book. Subsequently, we give a short summary of experience to date in measuring efficiency in the health sector. The chapter ends with an outline of the remainder of the book.

1.2 The demand for efficiency analysis in health care

The international explosion of interest in measuring the inputs, activities and outcomes of health systems can be attributed to heightened concerns with the costs of health care, increased demands for public accountability and improved capabilities for measuring performance (Smith 2002). Broadly speaking, the policy maker's notion of efficiency can be thought of as the extent to which objectives are achieved in relation to the resources consumed. There might also be some consideration of external circumstances that affect the ability of the system to achieve its objectives. This beguilingly simple notion of efficiency is analogous to the economist's concept of cost-effectiveness, or the accountant's concept of value for money. The potential customers for measures of efficiency include governments, regulators, health care purchasers, health care providers and the general public.

Governments clearly have an interest in assessing the efficiency of their health institutions. In all developed countries, public finance of one sort or another is the single most important source of health system funding, so national and local governments have a natural requirement to ensure that finance is deployed effectively. It is therefore not surprising to find that methodologies that offer insights into efficiency have attracted the interest of policy makers. Moreover, in most industrialised countries, a large element of the health care sector is provided by non-market organisations. Given the complexity of the functions undertaken by such institutions, and in the absence of the usual market signals, there is a clear need for instruments that offer insights into performance. The search for such technologies has been intensified by the almost universal concern with escalating health care costs and increased public pressure to ensure that expenditure on health systems is used effectively.

Given the absence of a competitive market in health care, all health systems require a regulator of some sort. A regulator is most obviously required when a significant proportion of health care is provided by the for-profit sector. However, the regulatory function might be incorporated implicitly into government surveillance of the health system if public provision predominates. As well as having an obvious role in promoting public safety, effective regulation requires the development of measures of comparative performance in order to set a level playing field for providers, a task to which efficiency models are in principle well-suited. Such interest is of course not limited to the health sector. For example, the UK water industry regulator (OFWAT) makes extensive use of efficiency analysis in determining its regulatory regime for water companies (Office of Water Services 1999).

Health care purchasers have a serious information difficulty when negotiating contracts with providers. In the absence of any meaningful market, they often find it difficult to judge whether providers are offering good value for money. Even in a competitive environment, it may be difficult for purchasers to discriminate between competing providers. Efficiency analysis can therefore help purchasers to understand better the performance of their local providers relative to best practice, and introduces an element of 'yardstick competition' into the purchasing function (Schleifer 1985). Likewise, even in non-competitive health care systems, providers have a natural interest in seeking out best practice and identifying scope for improvement.

Finally, there are increasing demands for offering the general public reliable information about the performance of its national and local health systems, and of individual providers (Atkinson 2005). Whilst the evidence hitherto suggests that it is difficult to stimulate public interest in this domain – and we are not aware of any major initiatives involving efficiency analysis – there are strong accountability arguments for seeking to place high-quality information in the public domain in order to enhance debates about value for money.

1.3 Organisational efficiency

The focus of efficiency analysis is as an organisational locus of production, often referred to as a decision-making unit (DMU). In health care, examples of DMUs include entire health systems, purchasing organisations, hospitals, physician practices and individual physicians. The DMUs consume various costly inputs (labour, capital etc.) and produce valued outputs. Efficiency analysis is centrally concerned with measuring the competence with which inputs are converted into valued outputs. In general, it treats the organisation as a black box, and does not seek to explain *why* it exhibits a particular level of efficiency (Fried, Lovell and Schmidt 1993).

The terms 'productivity' and 'efficiency' are often used interchangeably, which is unfortunate since they are not precisely the same thing. Productivity is the ratio of some (or all) valued outputs that an organisation produces to some (or all) inputs used in the production process. Thus the concept of productivity may embrace but is not confined to the notion of efficiency that is the topic of this book.

A starting point for examining the basic notion of efficiency is shown in Figure 1.1, which illustrates the case of just one input and one output. The line OC indicates the simplest of all technologies: no fixed costs and constant returns to scale. A technically efficient organisation would then produce somewhere on this line, which can be thought of as the production possibility frontier. Any element of inefficiency would result in an observation lying strictly below the line OC. For an inefficient organisation located at P_0, the ratio $X_0P_0/X_0P_0^*$ offers an indication of how far short of the production frontier it is falling, and therefore a measure of its efficiency level.

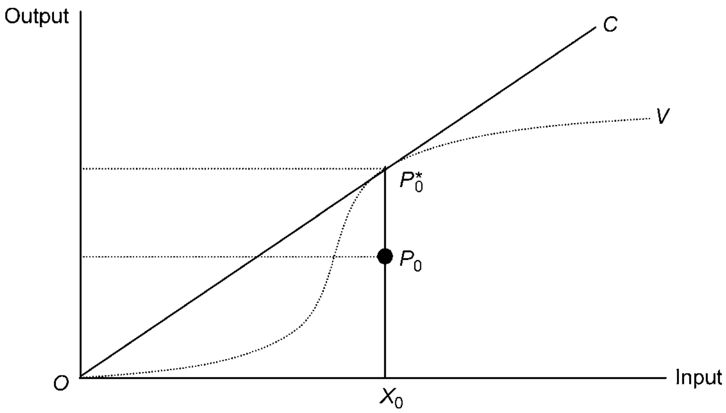

Figure 1.1. Efficiency measurement under constant returns to scale.

Many other technologies are possible. For example, the curve OV indicates a frontier with variable returns to scale. Up to the point P_0^*, the ratio of output to input decreases (increasing returns to scale), but thereafter it increases (decreasing returns to scale).

The notion of a production frontier can be extended to multiple outputs and a single input (say, costs). Figure 1.2 illustrates the case with two outputs. For the given technology, the isocost curve CC gives the feasible combination of outputs that can be secured for a given input. At a higher level of costs the isocost curve moves out to $C'C'$. These curves indicate the shape of the production possibility frontiers at given levels of input. An inefficient DMU lies inside this frontier. We define the marginal rate of transformation to be the sacrifice of output 2 required to produce a unit of output 1, indicated at any particular point on CC by the slope of the curve $-(P_2/P_1)$. It is usually assumed that – as in this figure – for a given level of input this becomes steeper as the volume of output 1 produced increases.

Likewise, in input space, we examine the case of two inputs and one output, as in Figure 1.3. The isoquant QQ indicates the feasible mix of inputs that can secure a given level of output, with inefficient DMUs lying beyond this curve.

Extending the analysis to the general case of multiple inputs and multiple outputs, we define the overall efficiency eff_0 of organisation 0 to be the ratio of a weighted sum of outputs to a weighted sum of inputs. Mathematically, if organisation 0 consumes a vector of

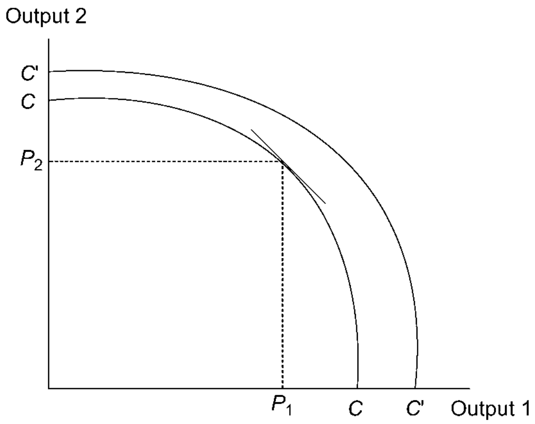

Figure 1.2. The case of two outputs.

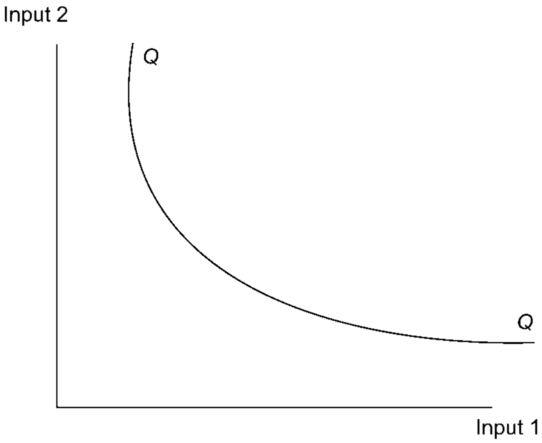

Figure 1.3. The case of two inputs.

M inputs $\mathbf{X}_0$ and produces a vector of S outputs $\mathbf{Y}_0$, its overall efficiency is measured by applying weight vectors $\mathbf{U}$ and $\mathbf{V}$ to yield:

$$eff_0 = \frac{\sum_{s=1}^{S} U_s Y_{s0}}{\sum_{m=1}^{M} V_m X_{m0}} \tag{1.1}$$

where:

Y_{s0} is the amount of the *s*th output produced by organisation 0;
U_s is the weight given to the *s*th output;
X_{m0} is the amount of the *m*th input consumed by organisation 0;
V_m is the weight given to the *m*th input.

The weights **U** and **V** indicate the relative importance of an additional unit of output or input. On the input side, the weights **V** might reflect the relative market prices of different inputs. It is often the case – with the notable exception of capital inputs – that these can be measured with some accuracy. Then, if the actual input costs incurred by organisation 0 are C_0, the ratio:

$$Ceff_0 = \frac{\sum_{m=1}^{M} V_m X_{m0}}{C_0} \tag{1.2}$$

indicates the extent to which the organisation is purchasing its chosen mix of inputs efficiently (that is, the extent to which it is purchasing its chosen inputs at lowest possible prices).

However, the organisation may not be using the correct mix of inputs. This can be illustrated using a simple two-input model. For some known production process, the isoquant *QQ* in Figure 1.4 shows the use of minimum inputs required to produce a unit of a single output. The points P_1 and P_2 lie on the isoquant and therefore – given the chosen mix of inputs – cannot produce more outputs.

When the unit costs of inputs are known, it is possible to examine the input price (or allocative) efficiency of the two units. Suppose the market prices are V_1^* and V_2^*. Then the cost-minimising point on the isoquant occurs where the slope is $-V_1^*/V_2^*$ (shown by the straight line *BB*). In Figure 1.4 this is the point P_1, which is input-price efficient. However, the point P_2 is not efficient with respect to prices, as a reduction in costs of $P_2P_2^*$ is possible. The price efficiency of P_2 is therefore given by the ratio OP_2^*/OP_2.

Analogous arguments can be deployed to examine the allocative efficiency of organisations in output space. Figure 1.5 illustrates the case where a single input is used to produce two outputs. If the relative values U_1 and U_2 of the outputs are known, and the production possibilities are given by the curve *CC*, then organisation P_1 is producing

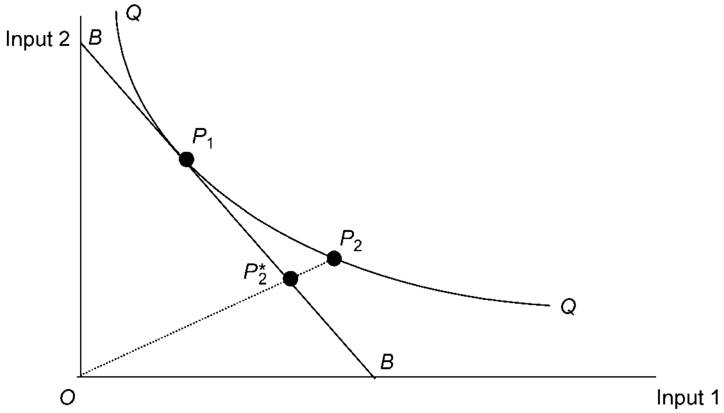

Figure 1.4. Allocative efficiency with two inputs.

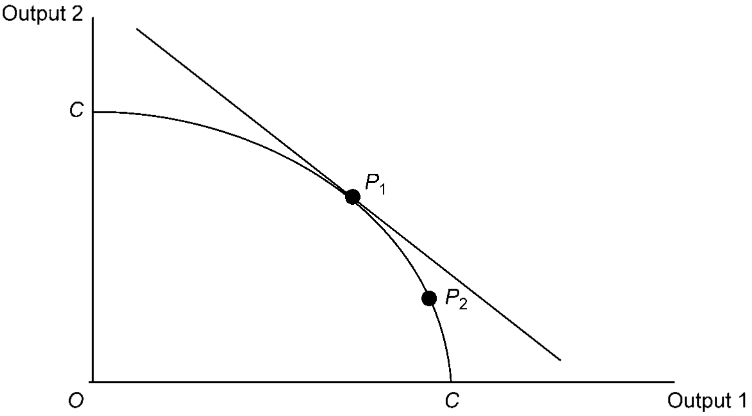

Figure 1.5. Allocative efficiency with two outputs.

at its allocatively efficient point while organisation P_2 exhibits some allocative inefficiency.

Although organisations may exhibit allocative inefficiency in purchasing the wrong mix of inputs or producing the wrong mix of outputs, we have so far explored only those organisations that lie on the frontier of technical production possibilities. However, it is likely that, particularly in a non-market environment, many organisations are not operating on the frontier. That is, they also exhibit an element

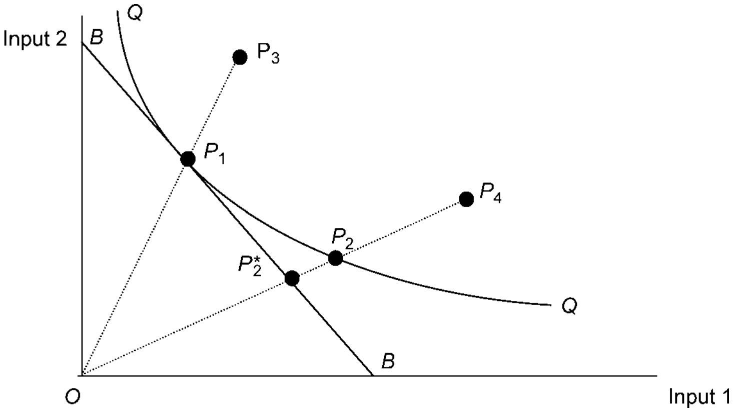

Figure 1.6. Technical and allocative inefficiency.

of technical inefficiency (also referred to as managerial inefficiency or X-inefficiency).

This is illustrated in Figure 1.6 by the points P_3 and P_4. Organisation P_3 purchases the correct mix of inputs, but lies inside the isoquant QQ. It therefore exhibits a degree of technical inefficiency, as indicated by the ratio OP_1/OP_3. Organisation P_4 both purchases an incorrect mix of inputs and lies inside the isoquant QQ. Its technical inefficiency is indicated by the ratio OP_2/OP_4. Thus its overall level of inefficiency OP_2^*/OP_4 can be thought of as the product of two components: technical inefficiency OP_2/OP_4 and allocative inefficiency OP_2^*/OP_2.

We have so far assumed constant returns to scale. That is, the production process is such that the optimal mix of inputs and outputs is independent of the scale of operation. In practice there exist important economies and diseconomies of scale in most production processes, so an important influence on eff_0 (from equation 1.1) may be the chosen scale of operation. This is illustrated in Figure 1.7 for the case of one input and one output. The production frontier is illustrated by the curve OV, which suggests regions of increasing and decreasing returns to scale. The optimal scale of production is at the point P^* where the ratio of output to input is maximised. Although lying on the frontier, the points P_1 and P_2 secure lower ratios because they are operating below and above (respectively) the scale-efficient point of production. They therefore exhibit levels of scale inefficiency given by:

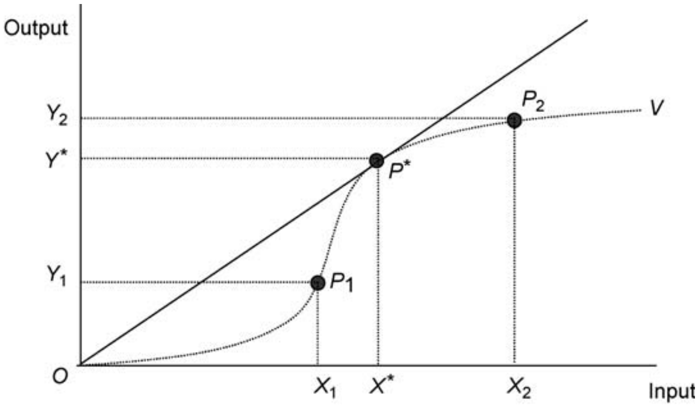

Figure 1.7. Economies of scale.

$$Seff_1 = \frac{OY_1/OX_1}{OY^*/OX^*} \text{and } Seff_2 = \frac{OY_2/OX_2}{OY^*/OX^*} \qquad (1.3)$$

1.4 Analytic efficiency measurement techniques

The fundamental building block of the economic analysis of organisational efficiency is the *cost function* (or its counterpart, the *production function*). For the purposes of this exposition, we focus on the cost function. This is probably more germane to the health care setting we seek to analyse, in which it is usual to find multiple outputs quantified on different measurement scales. The cost function simplifies the input side of the production process by deploying a single measure of the inputs used, rather than a vector. It indicates the minimum cost that an organisation can incur in seeking to produce a set of valued outputs. Using the notation introduced above, a cost function can be written in general terms as $C_0^* = f(\mathbf{Y}_0)$. Analogously, the production function models the maximum (single) output an organisation could secure, given its mix of inputs.

The cost function combines all inputs into a single metric (costs), and does not model the mix of inputs employed, or their prices. In practice, the costs incurred by an organisation might be higher than those implied by the cost function for three reasons. First, it may purchase inputs at higher than market prices (cost inefficiency). Second, given prevailing prices, it may employ an inefficient mix of inputs