COMMISSION ON MACROECONOMICS AND HEALTH
A WHO commission examining the interrelations among investments in health, economic growth and poverty reduction

CMH Working Paper Series

Paper No. WG3 : 2

Title

Financing Health Systems through Efficiency Gains

Author

M. Hensher

Date: May 2001

DRAFT

Financing Health Systems Through Efficiency Gains

Martin Hensher

Working Paper for the

Commission on Macroeconomics and Health

Working Group 3 – Resource Mobilization

6th May 2001

Acknowledgements

The production of this draft report would not have been possible without the enthusiasm and patience of my research assistants, Kisimba Mwenge and Etienne Yemek of the Department of Economics, University of Pretoria. I am also very grateful for the assistance and support of Vishal Brijlal, Director: Health Financing and Economics in the Department of Health. Most important, however, has been the support, good humour and understanding of my wife Andrea, who has not once asked why I was reading an obscure article when I should have been changing a nappy.

Disclaimer

The views and opinions expressed in this paper are those of the author in his personal capacity, and in no way reflect the policy or views of either the European Commission or the Government of the Republic of South Africa.

Contact Details:

Martin Hensher
European Union Consultant in Health Economics
Directorate: Health Financing & Economics
Department of Health
Private Bag X828
Pretoria
0001
South Africa

henshm@worldonline.co.za

Tel: +27 12 312 0672
Fax: +27 12 329 2588

Introduction

Background

In recent years, much of the analytical and intellectual effort concerned with the economics of health systems in developing countries has tended to focus on aspects of what will be defined in this paper as "allocative" efficiency – broadly, achieving optimality in the distribution of resources between competing uses. Specifically, much effort has been devoted to the development and contestation of techniques of cost-effectiveness analysis, which aim to allow comparison and choices between alternative health interventions and programmes (e.g. Ghana Health Assessment Project Team, 1981; World Bank, 1993). At the same time, considerable intellectual and policy effort (although perhaps rather less analysis) has been applied to attempts to expand or mimic the role of markets in health care. The consideration of "technical" efficiency (broadly, achieving lowest input / lowest cost production of any given output) has figured less prominently in the search for "solutions" to the problems of health systems in poor countries. In the best traditions of economics, those working on the economic evaluation of health care interventions and programmes have tended to adopt the assumption, be it explicitly or implicitly, that the interventions they are examining are being produced at least possible cost (just as classical theory assumes the firm will in all cases minimise costs). Indeed, this author has more than once heard the assertion that the difference in outcomes between competing health programmes will "almost always" dominate any difference in cost arising from different levels of technical efficiency. Clearly, though, failing to account for differing real levels of technical efficiency between providers or between health systems in different countries could have significant implications for the generalisability and validity of the results of economic evaluations, be they at intervention or sectoral level.

Recognising the potential impact of technical efficiency upon key policy choices, this paper was therefore commissioned to "…provide the Commission on Macroeconomics and Health and developing countries with a practical assessment of how to improve technical efficiency in order to free up resources to finance their health needs."

Structure of the Paper

The paper considers a range of efficiency concepts, and develops working definitions for application in the remainder of the paper, before considering some specific

characteristics of health care production which are of relevance. A typology of inefficiencies likely to be encountered in health care provision is then developed, and different causes of such inefficiencies are considered. The possible theoretical impacts of technical inefficiency (and improvements therein) upon questions of allocative efficiency are then discussed, with practical examples where possible. Alternative approaches (both academic and managerial) to the measurement of technical efficiency are then discussed, and their suitability for application in developing countries explored. The international literature on the existence of technical inefficiencies, and the potential scale of savings from their elimination is then reviewed.

A framework is then proposed for application at country level for the identification of significant inefficiencies, the development of strategies to improve efficiency in the production of health services, and for ensuring best fit between improvements in technical efficiency and other health sector and/or macroeconomic goals.

Efficiency in the Production of Health Care – Concepts and Definitions

There is a surprising range of definitions of certain core efficiency concepts on offer (see e.g. Liu, forthcoming). This section identifies the central concepts, briefly discusses any significant variations in their interpretation, and then develops working definitions to guide the remainder of the paper.

Production

Underlying all conceptions of productive efficiency is, of course, the notion of production itself – production being any process which converts or transforms a commodity (which may be a physical good or a service) into some other *different* commodity (e.g. Lancaster, 1969; Rice 1998). The process of production brings together different *inputs* (or *factors of production*), such as labour, land, capital in the form of buildings and machinery, and raw materials or intermediate inputs (an output produced by someone else which goes on to be used as an input in another production process), combines them using knowledge of a technical process, and thus creates an *output*. This process of combination is conventionally conceived of as the *production function* – the functional relationship by which inputs $I_a \ldots I_n$ are converted into output O. It is important to note that, even in the most resource-poor and technologically basic settings, the production of health care requires an extremely wide range of inputs. These include many different types of labour input (physicians and nurses with different specialist training, therapists, radiologists,

laboratory technicians, paramedics, administrative and clerical staff, cleaners, drivers - to name but a few), clinical and diagnostic equipment ranging from a stethoscope to a linear accelerator, drugs, laboratory reagents, consumables, linen, meals, vehicles, buildings and plant…

The outputs of health care have often been seen to pose a particular problem for economic analysis. It is generally accepted that the ultimate output of health care is improved "health" in the recipient patient (e.g. Williams, 1986) – but measuring "health status" remains heavily contested technically, methodologically fraught, very expensive and very hard to operationalise even in ideal research sites. Equally, health care systems provide services for patients which can certainly be seen as, at least, intermediate outputs – a coronary artery bypass graft, a course of antibiotics, a completed course of immunisation. In this paper, almost all discussion of health care outputs will, in fact, relate to such intermediate outputs of health care production processes, for reasons which will rapidly become obvious.

Technical Efficiency

A technically efficient firm or production unit produces as much output as possible with a given amount of inputs, or produces a given output with the minimum possible quantity of inputs. In terms of classical textbook exposition, a technically efficient firm produces on the isoquant / production possibility frontier, while a technically inefficient firm operates off the isoquant, or inside its production frontier (McGuire, 1987; Barnum and Kutzin, 1993). Critically, almost all mainstream definitions take technical efficiency to refer only to input quantities, and not input costs in monetary terms (for an exception, see Hurley *et al*, 1995). Several authors raise the related issue of technological efficiency (e.g. Lancaster, 1969). Technological change occurs through the development of new processes which can produce more output for the same or less input than older processes; they argue that the introduction of such a new process can be thought of as rendering all previous processes technically inefficient. Under this view, "technology' consists of the series of all known techniques for producing a particular output – although the invention of a new technique does not necessarily mean it will be available to all producers or all countries (Meier, 1995). Clearly, though, there is a difference between inefficiency due to operating off the isoquant for a given technology, as opposed to inefficiency due to failing to move to a different isoquant made possible by a new technology. An example of the former would be a situation in which an equivalent care outcome could be achieved by administering a lower dose of a drug (Palmer and Torgerson, 1999) – while the latter could be exemplified by the introduction of a new drug identify

two forms of technological change: *process innovation*, which improves production processes of existing products, and *product innovation*, which develops substantively improved outputs. Färe *et al* (1997) suggest that technological change represents *innovation* (substantively new processes or new products), while improving technical efficiency under a given technology is essentially about "catching up" to what is already possible (while they are discussing country-level performance, this conception seems equally applicable at lower levels).

Economic Efficiency

The cost of any production process is, of course, influenced not only by the quantities of inputs used, but also by the cost of these inputs. A production unit which is *economically efficient* will produce a given output for the minimum possible total input *cost*, or maximize output for a fixed value input budget. Thus, an economically efficient firm is, by definition, a cost-minimiser. Kleczowski (1980) and Gilson and Mills (1995) provide the same definition (i.e. least cost production of a given quantity / maximum production for a fixed budget), but under the name of *operational efficiency*. In terms of conventional exposition, achieving economic efficiency requires that a firm operate at the point of tangency between the isoquant and the isocost line. At this cost-minimising point, the value of the marginal product of each factor will be equal ($MP_K/P_K = MP_L/P_L$), also expressed as saying that the marginal rate of technical substitution between inputs is equal to the ratio of their input prices. When these ratios are unequal, the potential exists to make factor substitutions which will reduce the costs of production.

The critical implication of economically efficient behaviour is the "principle" or "rule" of substitution (e.g. Lipsey and Chrystal, 1995; Samuelson and Nordhaus, 1995) – namely, that when the relative prices of inputs change (for example, if the prices of imported drugs rise relative to the costs of labour due to exchange rate depreciation) the choice of production process will change to use relatively more of the cheaper factor and relatively less of the more expensive factor.

Technical efficiency can never exclusively determine economic efficiency, as the values of inputs can always change.

> "No machine, no process, no arrangement is so efficient that it cannot be rendered inefficient (or so inefficient that it cannot be made efficient) by an appropriate change in values." (Heyne, 1994)

This formulation of economic efficiency is particularly important in considering health care interventions. Clinicians (quite reasonably) tend frequently to focus on best

practice in terms of inputs – but differences in relative input prices may mean that a technically efficient "best practice" is economically efficient in one country but not in another. This possibility is clearly a key practical constraint upon attempts to produce truly international "evidence based medicine" and to develop easily generalisable cost-effectiveness results.

In defining efficiency in terms of outcomes, rather than outputs, Palmer and Torgerson (1999) introduce a broadly parallel concept which they call *productive efficiency*. This, they argue, can be used to assess the efficiency of alternative interventions which have directly comparable outcomes but different inputs, hence requiring comparison in terms of the cost of inputs (for example, surgical *vs* antacid *vs* eradication of *h. pylori* as treatment for stomach ulcers). This formulation is clearly equivalent to cost-minimisation analysis, and as such represents a particular case of economic evaluation (see Drummond *et al* 1997). The value and validity of cost-minimisation analysis for economic evaluation has recently been questioned by Briggs and O'Brien (2000), who argue that current techniques may fail to identify small but real differences in outcome.

Productivity

The production function for a particular process thus represents the relationship between outputs of goods and services in real physical ("primal") volumes to the different inputs used, also in terms of physical volumes, which can be expressed in terms of output per unit of total input (O/I) – or *productivity* (Kendrick *et al*, 1981). Productivity can be measured through the use of *partial* productivity measures, the ratio of output to an individual input or input class (e.g. output per worker, output per hour of labour; clinic visits per nurse per day, analyses per machine hour etc.), or in terms of *multi-factor* productivity (or total factor productivity), the ratio of output to all associated inputs. Changes in multi-factor productivity are directly equivalent to changes in the economic efficiency of production, in that they reflect improvements in the real cost of production over time (ABSSP, 1979). Measures of partial factor productivity are attractive because they avoid the need for monetary valuation of inputs and for the calculation of constant prices over time (Mahoney, 1980), and can be used to illustrate savings achieved over time (or variations between similar production units) in the use of particular inputs. However, they have the potential to mislead, as they reflect not only improvements in the productive efficiency of the input in question, but also changes in output which resulted from factor substitutions made in response to changes in relative factor prices. Thus, for example, the introduction of digital film-free X-ray imaging may lead to an increase in X-rays per

radiographer – but this productivity increase is not due to any effort on the part of radiographers, but to the substitution of capital equipment for what was previously a manual task (film development). It is also important to note in passing the important tendency towards a *diminishing marginal rate of substitution* between inputs. This describes the situation in which, as we increasingly substitute one input for another, the less effective will the former become as a substitute for the other. Thus, in a situation where we were increasingly substituting labour for capital in a production process, we would find ourselves having to substitute progressively more worker hours for each machine-hour we give up (hence explaining why some capital is required in almost any production process, no matter how cheap labour may be in comparison).

Allocative Efficiency

In contrast to the technical and economic efficiency concepts discussed above, which all consider only the process of production, concepts of *allocative* efficiency embrace the notion that society is concerned not just with how an output is produced, but also with what outputs and what balance of outputs are to be produced. Thus allocative efficiency is conventionally defined as being achieved in a situation in which it is impossible to improve the welfare of anyone without reducing the welfare of someone else through a change in the output combination (Lindsay, 1982; Lipsey and Chrystal, 1995); i.e. the achievement of a Pareto-optimal state (Rice, 1999). Palmer and Torgerson (1999) describe this concept as the efficiency with which these outcomes are distributed among the community, and Donaldson (1994) goes on to suggest that what separates notions of allocative efficiency from technical efficiency is that the former is concerned with the question of *who* benefits from production, while the latter concerns only production itself. Explicitly, technical and economic efficiency are necessary but not sufficient conditions for allocative efficiency to be achieved. Knox Lovell and Schmidt (1988) present a neat summary of what this entails for the individual firm:

> "It [the firm]…produces the correct mix of outputs, given output prices, uses the correct mix of inputs, given input prices, and adopts the correct scale given input and output prices: this is what allocative efficiency requires."

Considering the health sector specifically, Kleczowski (1980) suggests that there are two dimensions of interest in allocative efficiency terms: a wider sense, which concerns the overall adequacy of investment and production of the health sector relative to society's wider welfare, and a narrower sense, which concerns the balance

of resource distribution within the health sector. In recent years, a common usage of the term allocative efficiency has been adopted in health care which refers increasingly to the idea that society's health status should be maximised, through achieving the most cost-effective balance of programmes and interventions. Through this usage, sectoral cost-effectiveness analysis (e.g. through the use of DALYs etc.), cost-utility or cost-benefit analysis can be seen as providing information on allocative (in)efficiency in health care.

Clearly, considerations of allocative efficiency and inefficiency, especially under conditions of market failure, form the core business of welfare economics. The pre-eminent focus of both empirical and theoretical health economics on this subject is therefore perhaps not surprising, as this is probably the strongest point of linkage with "mainstream" economic theory.

Interestingly, one contemporary health economics textbook (Folland *et al*, 1997) defines allocative efficiency for the firm simply as requiring production at the point of tangency of the isoquant and the isocost curve – i.e. the general definition of *economic* efficiency. This conception of allocative efficiency also appeared in two of the applied papers reviewed (Rosko and Chilingerian, 1999; McGuire, 1987).

Short Run and Long Run

The concept of the *short-run* and the *long-run* have an essential role to play in considering efficient production and the extent to which inefficiencies can be reduced or eliminated. These concepts concern the extent to which, over time, a production unit can change the level and combination of inputs it employs, and/or the level or type of output it produces. Conventionally, the *long-run* refers to a period which is sufficiently long for a production unit to be completely free in its decisions from its present policies, possessions or commitments (Baumol, 1977). In other words, in the long-run, all current staff could be disposed of and an entirely new workforce could be employed, the current physical infrastructure demolished and replaced with completely different capital assets, etc. In contrast, in the *short-run*, at least one significant factor of production cannot be changed, i.e. is *fixed*. Short-run constraints can typically be influenced by the duration of contracts entered into (e.g. employment contracts, supply contracts etc.) which may be impossible or costly to break; lack of availability of appropriate skills, and the pipeline time required to obtain these skills through additional training, recruitment, migration etc.; the lead time required to initiate and build or convert new facilities etc – or through unavailability of funds with which to employ a necessary new input. Complete long-run adjustment of inputs

clearly is likely to require substantial periods of time, and many production units may find themselves constrained to some extent by previous commitments for many years. These constraints mean that, in practical terms, many producers may struggle to reach technically or economically efficient production positions. Feasible adjustment times (the movement from the short- to the long-run) therefore have significant implications for improving efficiency through constraints on effective resource substitution within processes due to factor availability or mobility, and on the ability to transform fixed inputs back into cash (a frequently overlooked but necessary condition for transferring resources from one programme into an unrelated use that does not share the same physical input requirements).

Returns to Scale

Also essential to any consideration of productive efficiency are the concepts of returns to scale and scope. A firm is said to exhibit *constant* returns to scale when a unit increase in inputs yields a proportionate unit increase in output. *Increasing* returns occur if a unit increase in input yields a proportionately larger increase in output, and *decreasing* returns when a unit increase in input yields a proportionately smaller increase in output. Equivalently, a firm is said to exhibit *economies (diseconomies) of scale* if, over some portion or all of its long-run average cost curve, average costs are declining (increasing). Returns to scale are a long-run concept, and should not be confused with short-run reduction of average cost due to increased utilisation of fixed inputs. "Real" economies of scale are held to derive purely from technical indivisibilities in production (i.e. you can't have half an operating theatre), while "pecuniary" economies of scale can be yielded by larger organisations through their greater purchasing power (e.g. through the ability to negotiate bulk quantity discounts on supplies, or monopsony wage power in specialist labour markets). Where economies of scale can be shown to exist, they may represent an important factor to be considered in efforts to improve the efficiency of health systems. *Economies of scope* may occur whenever it is possible to produce two or more outputs jointly more cheaply than they can be produced separately. They may occur due to the ability to share indivisible resources which would generally never be fully utilised at any realistic scale of single output production. Key practical examples in health care might include the sharing by several specialties of relatively large and expensive items of diagnostic equipment (e.g. an MRI scanner), which would not be fully utilised in a single specialty hospital; or specialists providing both inpatient and ambulatory care on a single hospital site. Economies of scope might also be generated by synergistic interaction between disciplines, leading to innovation.

Working Definitions and Scope

Given the foregoing discussion, it is possible to provide the following working definitions with which to guide the scope of this paper. The analysis and discussion will focus on the technical *and* economic efficiency of health services, which we take to mean achievement of (or deviation from):

- Technical efficiency i.e. maximising output for a given set of physical inputs / minimising the physical inputs required to produce a given output

  and

- Economic efficiency i.e. maximising output for a given budget / minimising production costs for a given level of ouput

We shall take allocative efficiency to mean the achievement a pareto-optimal distribution of health benefits across society which, insofar as it differs from technical and economic efficiency, is concerned practically with the optimal balance of production across health programmes.

Some Health Sector Specific Issues

Before concluding this conceptual introduction, it is important to highlight a number of issues which, while not unique to health, are of particular import in considering "efficiency" in the health sector. Almost all health care programmes (with the possible exception of highly targeted vertical programmes such as child immunisation campaigns) produce multiple outputs and have multiple goals. The production of multiple outputs poses challenges for measurement and comparability, which are not insuperable, but may be significant in the context of developing countries where data systems and personnel with the skills to interpret data appropriately may be scarce. To compound issues, these outputs may be highly heterogeneous with respect to quality – operator A may achieve quite different results from operator B for the same surgical procedure, equipment and patient types, while patient compliance with treatment can significantly affect outcome. Simplification is therefore unavoidable, and frankly nothing to be ashamed of, especially when, as in the case of this paper, the level of analysis is necessarily highly aggregated and seeking to achieve maximum generalisability. The *de facto* acceptance of the necessity to consider productive efficiency in terms of intermediate outputs, rather than changes in health status, has already been alluded to. Nevertheless, even comparison of apparently similar intermediate outputs (e.g. primary care visits, hospital admissions etc.) must be thoroughly hedged with caveats – the comparability of similar studies of similar

organisations or systems may well be significantly affected by underlying differences in population characteristics, case-mix and service organisation.

It should already be clear that two further conceptual problems are likely to make comparison and generalisation from one country to another especially difficult in terms of technical and economic efficiency. As the preceding discussion indicated, there can be no absolute measure of economic efficiency, as this is determined by relative factor prices. Hence, what constitutes an economically efficient process in one country may be inefficient in another, due to different relative factor prices. Attempts have been mooted to attempt to tackle this problem in cost-effectiveness analysis through devising suitable weighting measures (e.g. the use of health sector purchasing power parities), but these have yet to prove their worth in this task. Furthermore, even within a country, productivity and efficiency (of all forms) are essentially relative concepts. There is no absolute "ideal" level of productivity or technical efficiency, as technology and processes are constantly changing and evolving, and we have already discussed the relative nature of economic efficiency. Therefore, comparators and benchmarks can never be more than guides towards improvement and what might be possible, and we can be quite certain that no health system (or any other industry, for that matter) will ever be perfectly efficient in any of the dimensions of efficiency we have identified.

Typology and Examples of Technical and Economic Inefficiencies in Health

Before moving on to examine the available empirical literature, it is worthwhile to give a little more attention to concepts of inefficiency, and to identify the form in which such inefficiencies might be most likely to manifest themselves in health services. The discussion of efficiency concepts has implicitly identified the four main conceptual sources of technical and economic inefficiency (which may, of course, occur independently or in parallel). These are:

- Failing to minimise the physical inputs used (i.e. operating within the production possibility frontier)
- Failing to use the least cost combination of inputs (i.e. failing to operate at the point of tangency between the isocost curve and the isoquant)
- Operating at the wrong point on the short-run average cost curve
- Operating at the wrong point on the long-run average cost curve

Fairly obviously, the first two forms of inefficiency cover a multitude of sins, while the latter two are much more specific cases (although nonetheless significant). Examples can be manufactured almost indefinitely, but it is useful to illustrate some

of the more important and commonly encountered variants, to provide an anchoring point for the discussion which follows.

Failing to minimise the physical inputs used

- Excessive hospital length of stay, with patients remaining in hospital after they have ceased to benefit from hospitalisation
- Poor scheduling of diagnostics and procedures, resulting in excessive hospital stay
- Prescribing an intervention or diagnostic test which is known to be of no therapeutic value or relevance
- Over-prescribing of drugs (too high a dosage, too long a course, more substances than are actually required)
- Excessive use of diagnostic tests (e.g. performing daily tests when the specialist will only be available to interpret them once a week)
- Wastage of stocks – allowing stocks to expire, or allowing deterioration due to poor storage etc.; discarding unused contents of opened packets
- Over-staffing

Failing to use the least cost combination of inputs

- Inappropriate overuse of more expensive staff relative to less expensive staff, e.g. physicians vs. professional nurses for basic prescribing of essential drugs, professional nurses vs. nursing assistants for basic personal care, professional nurses vs. clerical staff for basic administrative duties
- Use of branded drugs when generics are available
- Failure to secure lowest cost supply e.g. continuing to buy supplies from retail suppliers instead of through competitive bidding
- Being “locked in“ to purchasing consumables at a set price from a manufacturer for a piece of equipment which has been provided “free” or on loan
- Using paramedic-staffed emergency ambulances to transport patients home from hospital, instead of paying for their bus ticket

Operating at the wrong point on the short-run average cost curve

- Implementing budget cuts which protect salaries at the expense of other expenditure items, hence reducing the number of patients who can be treated, but with no reduction in fixed costs

- Refusing to fill a vacant anaesthetist post due to budget restraints, forcing the surgical staff to limit their operating time
- A rural hospital operating at an average bed occupancy of 50% due to limited local demand
- Inadequate drug supply leading to under-utilisation of primary care clinics

Operating at the wrong point on the long-run average cost curve

- Planning to provide full pathology laboratory facilities at every hospital when laboratory services actually demonstrate economies of scale
- Planning to build a 1500 bed teaching hospital when diseconomies of scale are known to operate in hospitals above 600 beds

The Causes of Technical and Economic Inefficiency

At base, there are two main reasons why firms or individuals might fail to minimise inputs and input costs. One explanation is that they are in fact seeking to minimise costs, but are being prevented from doing so – either by the sort of institutional constraints discussed in terms of the short-run cost curve above, or by information problems which prevent them from identifying efficient input combinations and processes. The other is that they are simply not trying to minimise costs, for some behavioural or motivational reason. We will return to the question of institutional and informational constraints later in this section, but it is important to deal first with the more fundamental question of whether the assumption that producers and individuals even try to minimise costs is reasonable or not.

Theoretical Background

The classical model of the profit-maximising firm clearly treated cost-minimisation as axiomatic – profits cannot be maximised if costs are not minimised. Yet, as the twentieth century progressed, evidence began to grow that many private firms demonstrated behaviours which appeared neither to maximise profits nor to minimise costs. At the same time, the extension of economic theory and techniques into non-profit areas of the economy begged the question of what sort of objectives might be followed if the goal of profit maximisation was not relevant. Two groups of theories were developed in response to these perceived problems with classical theory (Lee and Mills (1985) provide a useful summary). The first grouping took it as axiomatic that firms or individuals (or managers, or workers) were attempting to maximise *something* (described as the objective function) which provided them with utility.

Thus Baumol (1977) suggested that managers may attempt to maximise sales revenue (perhaps subject to a minimum profit constraint to keep shareholders happy), and Williamson (1963) suggested that they are seeking to maximise managerial discretion (to pursue their own utility). The second grouping took a bolder step, in suggesting that firms and individuals may not necessarily display maximising behaviour. In particular, "satisficing" theories posited that individuals and organisations attempt in the main to meet minimum acceptable standards or targets (e.g. Simon, 1959; Cyert and March, 1963) – perhaps an acceptable level of profit or market share, or effort / activity. All of these approaches clearly can explain deviation from cost-minimising behaviour, but approach the issue rather indirectly. In 1966, Leibenstein addressed the issue head-on, with the development of his theory of "X-efficiency", which suggested that firms may fail to produce on their production possibility frontier and fail to minimise costs due to a combination of imperfect knowledge of production functions, imperfect employment contracts which create "slack" and discretion for workers and managers, and imperfect factor markets for managerial knowledge, which limit diffusion of managerial skills. Critically, the theory suggests that organisations operate with some degree of slack, and tend towards inert areas (i.e. where little change to process occurs) when they are not under pressure. The theory of X-efficiency generated significant implications for the potential role (and limits) of incentives, and of the likely importance of either competitive or regulatory pressure to encourage cost-minimisation. These departures from maximising theories were bitterly contested by some:

> "Unless one is prepared to take the mighty methodological leap into the unknown that a non-maximizing theory requires, waste is not a useful economic concept. Waste is error within the framework of modern economic analysis, and it will not become a useful concept until we have a theory of error." (Stigler, 1976)

Ultimately, though, while failing to provide an equilibrating model (their main crime against traditional economic thought), the behavioural theories seem to provide rather more useful explanatory capabilities than the catch-all of "utility maximisation", and are applicable in non-profit situations – a critical consideration given the importance of public sector and non-profit health care provision in most countries.

Factors Predisposing Towards Technical and Economic Inefficiency

Somanathan *et al* (2000) provide an excellent review of the literature on the institutional features of public sector provision of health care and their impacts on

efficiency. They distinguish between *absence of incentives* for efficient behaviour, and *constraints* on decision-makers' abilities to make efficient choices. Table 1 adapts and extends their discussion of constraints and incentives which may predispose health systems to produce care inefficiently, and attempts to describe possible manifestations of each constraint / incentive at three levels: *micro*, meaning the level of individual health care workers or provider units; *system*, meaning the health care delivery system or sector as a whole; and *macro*, meaning the macroeconomic, society or government level.

Table 1 – Factors Predisposing Towards Technical and Economic Inefficiency (Adapted from Somanathan *et al*, 2000)

INCENTIVES		**Impacts by level:**		
Factor	**Key Features**	**Micro**	**System**	**Macro**
Public Ownership	No claim on residual profits / savings	Reduced incentive to minimise costs	Unable to retain savings within sector, therefore no incentive to minimise costs below level necessary to keep in budget	System seen not as an asset but as an expenditure liability, leads to under-investment
Objectives	Multiple policy objectives, requiring trade-offs and discretion (e.g. efficiency, equity, access, participation, quality, cost-recovery etc.)	Probably poorly understood and often contradictory at delivery level, leading to unclear incentives, confusion, satisificing responses and "doing what we've always done"	Inevitable trade-offs mean that no single objective is likely to be maximisable, including cost-minimisation	Non-health macroeconomic policy objectives may further constrain ability to operate efficiently, e.g. expenditure cuts, duties on imports, refusal to allow retrenchment of public sector workers etc.
Payment Mechanism - Personnel	Salaried systems will generally mean remuneration is unaffected by performance	Salaries provide little incentive to improve productivity, incentivised schemes may focus (distort) effort towards specified areas, often with unpredictable results	Salary systems good for cost control, despite lack of incentives. Performance related pay systems can be unfair and administratively complex, and with mixed evidence of success	Often difficult to reconcile centralised pay bargaining with performance based systems; nationally negotiated and unfunded or underfunded pay settlements can undermine efficient local operation
Payment Mechanism - Provider	Fee-for-service and case-based payments provide strong incentives to increase activity / revenue compared to fixed budgets	Incentives under FFS to maximise revenue may undermine cost-minimisation, and lead to inappropriate care and supplier-induced demand	Introducing payment incentives within a fixed budget may have very unpredictable impacts and distort priorities; without a hard budget, they will lead to cost-escalation	Setting up cost escalation may divert resources from higher priority uses

INCENTIVES (cont'd)		**Impacts by level:**		
Factor	**Key Features**	**Micro**	**System**	**Macro**
Market Information	No competitive performance information / signals	No comparative benchmarks, lack of pressure for improvement or innovation, leading to X-inefficiency	Little possibility to exercise judgements on performance and to reward / punish	Lack of information on outputs or performance leads to a sense that health sector is an "unproductive" black hole, leading to reluctance to invest additional resources
Corruption, theft and fraud	System tolerates corruption, theft and misappropriation of resources	Theft and wastage of resources increases costs and deprives patients of care; comes to be seen as compensation for inadequate salaries	Increased costs; misallocation of resources away from priorities towards areas which maximise rents; corrosion of management authority and accountability	Increased costs; misallocation of resources away from priorities towards areas which maximise rents; diversion of resources from poor to rich; corrosion of systemic integrity
CONSTRAINTS				
Lack of resources	Inadequate funding	Poor motivation due to low salaries. General resource shortages constrain ability to choose efficient input mix and undermine quality	Systematic tendency to skimp on "discretionary" expenditures (e.g. maintenance, training) leading to deterioration and increasing technical inefficiency over time. Donor dependence and resultant use of inappropriate technology	General lack of confidence in health system; health careers appear unattractive, leading to skill shortages. Increasing under-investment in human and physical capital in favour of recurrent items. Donor dependence
Input Indivisibilities	Under-utilisation of fixed, indivisible assets	Under-utilisation due to demand / supply mis-match leads to high fixed costs	Exacerbated by inflexible management procedures; tendency to try to increase utilisation ("good money after bad") rather than rationalising assets	Perceived under-utilisation may lead to a reluctance to increase funding or investment

CONSTRAINTS (cont'd)		**Impacts by level:**		
Factor	**Key Features**	**Micro**	**System**	**Macro**
Demand	Sparse rural population Preferences for inefficient and unnecessary care	Under-utilisation and high fixed costs; by-passing of lower levels of care; demand for technically inefficient care (e.g. excess or unnecessary prescribing)	High fixed costs due to dispersed population; private providers exploit demand for inappropriate care	Service duplication between public and private sectors, distortion of health priorities and often sub-standard care by private providers
Management Information Systems	Limited data availability on services for decision-makers, little data on prices for lower-level managers	No information on prices of many inputs and no involvement in procurement processes, hence little incentive or ability to minimise costs	Inadequate data to undertake performance management or to inform planning. Little diffusion of price information. No data or control in key areas (e.g. maintenance and transport) for which other sectors responsible	Absence of data with which to demonstrate health benefits, productivity and efficiency improvement reinforces stereotype of health as 'unproductive'
Public Sector Procedures and Policies	Conformity with centrally determined personnel, procurement and budgeting procedures	Limited local ability to change input mix due to e.g. employment contracts and policies, centralised procurement and supply, inflexible budgetary rules; limited local decision authority or financial delegation	Excessive centralisation of decision responsibility stifles flexibility; 'one size fits all' policies perpetuate inefficient input mixes; lack of integration in decision-making and procurement reduces opportunities to improve efficiency	Uniform employment and budgetary systems for different sectors may be appropriate for none; civil service employment procedures remove employment decisions and policy from provider units and even from sectors
Continued Dominance by Medical Profession	Physicians remain pre-eminent in health care management and decision-making	Resistance to and lack of professional management at local level; little pressure on individual physicians to justify service quality or resource use	Low status of medical administration fails to attract best candidates; lack of willingness to challenge clinical colleagues and resistance to introduction of performance management	Maintaining mystique of health's 'special' status hides poor performance. Ministers and civil servants who are clinically trained tend to lack professional political and managerial skills

Table 1 is certainly not exhaustive in its scope, but aims to identify key focal points around which inefficiency can breed. It serves to emphasise the point made by Parker and Newbrander (1994), namely that decisions and failings at every level of the system – from the consumer, through the individual health care worker and managers up to macroeconomic policy experts and politicians – can, with the best will in the world, encourage inefficient resource use. It particularly highlights one of the great unresolved problems of public sector management in developing countries – that successful public expenditure control can sometimes actually be the cause of technical and economic inefficiency, by constraining managers' ability to choose efficient resource mixes or operating levels.

It is appropriate to pick out the area of human resources management and remuneration for some additional comment, as it appears repeatedly in Table 1. It is a cliché, but nonetheless true, to say that, given that total personnel costs generally contribute between 60 to 80% of health care costs in practically all systems (e.g. Chernichovsky, 1980), any approach to efficiency improvement that fails to start with staffing is probably on a fast track to nowhere. Human resources policy therefore has the potential to be either an important support to, or major brake upon efforts to improve efficiency. Firstly, as far as is reasonably consistent with the prevailing social and economic praxis of the society in question, employment policy should support and facilitate efficient substitution of lower for higher cost labour inputs. Pursuit of a perfectly flexible piece-work market of itinerant doctors and nurses is hardly a feasible or a desirable objective. However, employment contracts should make some provision for reassignment of duties or redeployment (functionally or geographically), even if fixed-term contracts are not felt to be feasible (the latter clearly offer the opportunity of non-renewal, greatly facilitating skill substitution). Careful assessment of skill requirements and skill-mix needs to be undertaken regularly, so that opportunities presented by routine departures of staff (promotion, job moves, retirement etc.) can be exploited to allow skill substitution (do you really need to replace the retiring professional radiographer, or might not a practical radiography assistant be sufficient?). Clearly, this type of substitution can be a long-term process, requiring some patience and persistence – a new cadre of workers (even one requiring only a limited duration training input) can take several years to come on line in significant numbers. It is particularly important to remember that many past successes in introducing "new" classes of health workers in developing countries (e.g. village health workers) involved the introduction of a new service where none had existed previously, in an environment of general growth. This is

clearly different from (and easier than) achieving a reallocation of activities and responsibilities in an established operational environment – but this latter case is what factor substitution is actually about in most settings. Institutional and professional inflexibilities can easily sink attempts at skill substitution – or, even worse, lead to a situation where the new workers join the payroll, but none of the old group leave, potentially leading to even greater inefficiencies than before.

Remuneration policies and practices also seem likely to have a significant impact on the efficiency of service delivery in developing countries. Ensuring that the remuneration of skilled health workers is adequate seems frequently to be overlooked in the attempt to control costs and expenditure. However, there are several persuasive arguments as to why inadequate remuneration of skilled health workers will undermine efficiency. The first involves the generic theory of *efficiency wages* (e.g. Stiglitz, 1987; Yellen, 1984), which argues that productivity is directly affected by wage levels through attracting and retaining higher quality workers (who will stay with a firm as they will not be confident of matching this wage were they to leave), and through motivating higher levels of effort and morale, so that labour costs (in terms of cost per output) are actually minimised by paying a wage above the market-clearing rate. This argument may not hold when, in effect, the government sets the market-clearing rate for health workers – but where unemployment and/or alternative employment is a possibility, the general principle is likely to hold, i.e. that screwing wages down as low as they will go may actually be self-defeating. Second, adequate remuneration is required to ensure that high-quality individuals will not exit to the private sector or overseas (a persistent problem which punishes especially those countries who maintain high standards of training), and to direct them to the postings and assignments where they are most needed. Developing country public sector employment systems often lack the flexibility required to, for example, match pay offers or to offer enhanced non-pay benefits in order to retain key staff who are being poached – while benefit packages often reward staff for the "higher living costs" of staying in the capital city, rather than providing them with incentives to accept postings in deep rural areas. Persistently low pay will ultimately lead to particular professions becoming unattractive to new entrants (especially where entry requires university / tertiary level training), and hence to skill shortages. Overall, failure to attract and retain adequate quantities and quality of staff will lead to technical inefficiency because of skill shortages – unfilled vacancies in key posts mean critical activities do not take place, and efficient operation becomes significantly degraded. Finally, persistently low pay almost always opens the door to

unofficial "private practice" using public facilities and time, if not to full-blown theft and corruption. Once again, the wastage of resources and low productivity that result may well outweigh the "saving" in salaries achieved by a low wage policy.

Interactions Between Allocative and Technical / Economic Efficiency

As discussed earlier, it is not the objective of this paper to consider the scope for or methods of achieving improvements in the allocative efficiency of health systems (for a comprehensive review of the latter, see Liu, forthcoming). As a result, a number of exclusions are possible. Most critically, we will not be directly concerned with consideration of what choices of intervention are most cost-effective / health-maximising; or whether private or market solutions have any intrinsic efficiency advantage over public solutions; or with direct distributional consequences of different policy or service delivery options. There are, however, a number of potential interactions between improved (or decreased) technical and economic efficiency which do require theoretical discussion before proceeding to review the empirical results.

Technical Efficiency as a Precondition for Allocative Efficiency

First, and not to be under-rated despite its undergraduate tone, technical and economic efficiency are necessary conditions for the achievement (or improvement) of allocative efficiency.

> "Allocative efficiency infers operational efficiency: if something is deemed worth doing then it must be carried out in a way which ensures the optimum use of scarce resources." (Gilson and Mills, 1995)

An exclusive focus on, say, switching resources from less cost-effective to more cost-effective activities, will not realise its full benefits in terms of improved allocative efficiency if providers on the ground are not producing services at lowest cost (see Berman 1982 for a discussion of how operational factors may impact on "ideal" cost-effectiveness ratios). Failure to adequately incorporate some assessment of the relative efficiency of providers may therefore also bias the outcomes of cost-effectiveness analysis (e.g. a new intervention provided by a highly motivated and efficient provider is compared with standard care at a low-efficiency provider), or overstate the benefits achieved relative to "real-world" implementation. As a minimum, therefore, good practice in economic evaluation should seek to compare interventions between providers with similar operational efficiency levels, while

sensitivity analysis should attempt to consider the impact of different levels of technical and economic efficiency on results.

Allocative Consequences of Improvements in Techncial & Economic Efficiency

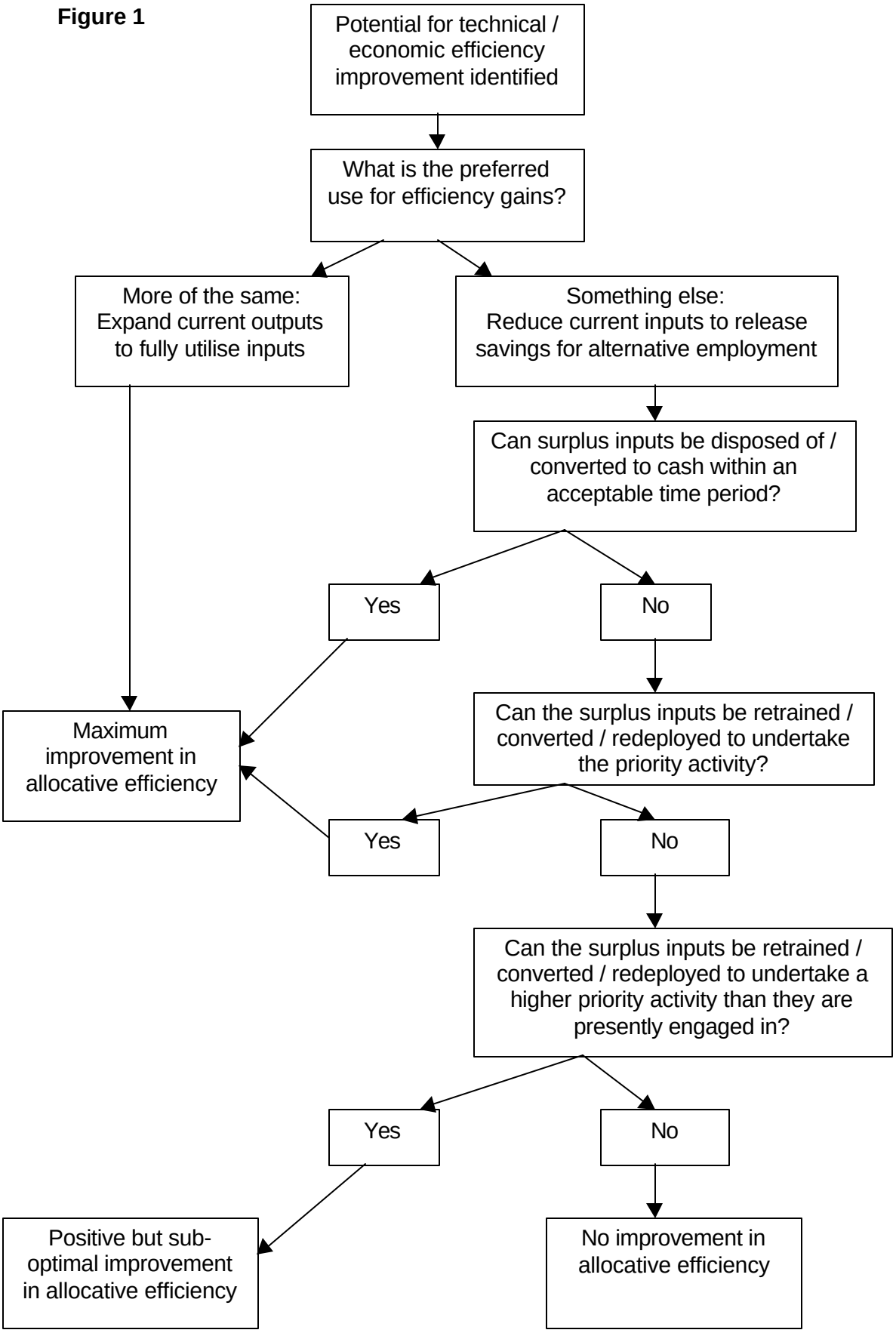
The second interaction requires rather more discussion, given the objective of this paper. It can be stated as follows. Improvements in technical and economic efficiency will release physical inputs or funds which were formerly tied up in inefficient production. This leads to i) a set of allocative decisions regarding what use these released resources should be put to, and ii) where, in the short-run, the resources released cannot be transformed into cash, potentially significant physical and technical constraints on the allocative choices available, and iii) a set of managerial decisions as to how practically to move from a short-run constrained position to the preferred long-run allocation of resources.

To develop this discussion, let us imagine two inefficient production units – one is a rural health centre, the other the department of nuclear isotope medicine in the national referral hospital. New management teams identify significant operating inefficiencies at both sites, in both cases relating primarily to low utilisation and high fixed staffing costs. Immediately, the nature of technical and economic efficiency means we face an allocative choice – do we want to maintain current output, and thus release inputs for other uses, or do we want to increase outputs until current inputs are efficiently employed? Do we want a higher utilisation of primary care in the community served by the health centre, or would it be more cost-effective (i.e. greater health gain per dollar spent) to take the efficiency savings and invest them into another programme, or another community? Do we want more patients to receive nuclear medicine diagnostic or therapeutic interventions, or would it be more cost-effective to take these efficiency savings and invest them into other programmes (e.g. primary care, immunisation etc.)? If, after careful consideration, we decide that it is indeed cost-effective and desirable to increase output of the existing providers and their services, then we are essentially faced with a series of tasks relating to improving the productivity of the current teams. If, however, we require only current output levels, and what we really want are the efficiency savings, then we face two sets of tasks – how to improve productivity (of those who are going to keep their jobs) *and* how to identify the surplus inputs and extract them / convert them into savings. It does not require great insight into human nature to realise that the second path is rather more arduous than the first for all concerned.

Having decided that we wish to put efficiency savings to some alternative use, we must then consider how far and how rapidly the current inputs can be converted to a new application – either physically (if they are suitable for redeployment), or via realisation into cash savings. This is clearly very much a problem of short-run versus long-run, and will largely depend on the extent to which institutional factors constrain the adjustment process. Returning to our example, let us imagine that our highest priority use for the efficiency savings is to establish primary health care clinics in currently under-served regions. Thus, at our inefficient health centre, we might have good reason to believe that some or all of the resources released by efficiency improvements could be directly transferred – for example, the staff probably already have the appropriate training, spare equipment could be moved etc. Even in this relatively straightforward case, we may still face constraints – will the redeployed staff be going to work for a new employer (perhaps a different regional health authority), and how will their contracts be transferred? Do their employment contracts allow us to transfer staff involuntarily? What period of consultation with unions may be required before definite decisions can be taken? The constraints are even stronger and starker in the case of our nuclear medicine department. Here, the surplus staff do not currently have appropriate skills to be transferred to frontline primary care – they are highly specialised radiographers, medical physicists and radiologists. Do procedures exist to allow us to initiate a process of redundancies? Do we have an effective human resources policy to allow us to select those who will stay and those who will go? If not, can we plausibly retrain the staff – if not to go to our highest priority programme, then at least to do something deemed more valuable than their current role (perhaps to work in X-ray departments in rural hospitals where there are skills shortages)? This option reduces the scale of the efficiency saving we will ultimately reap, but at least provides a solution that is less allocatively inefficient than the present situation. How long will all these processes take to work through?

The key allocative choices arising from improvements in technical and economic efficiency are summarised diagrammatically in figure 1 overleaf.

Figure 1


Potential for technical / economic efficiency improvement identified
What is the preferred use for efficiency gains?
More of the same: Expand current outputs to fully utilise inputs
Something else: Reduce current inputs to release savings for alternative employment
Can surplus inputs be disposed of / converted to cash within an acceptable time period?
Yes
No
Maximum improvement in allocative efficiency
Can the surplus inputs be retrained / converted / redeployed to undertake the priority activity?
Yes
No
Can the surplus inputs be retrained / converted / redeployed to undertake a higher priority activity than they are presently engaged in?
Yes
No
Positive but sub-optimal improvement in allocative efficiency
No improvement in allocative efficiency

Efficient and Appropriate Technology

It is also necessary at this point to touch briefly on the question of what constitutes "appropriate" health care technology. Conventional development economics identified and developed the concept of "appropriate" and "inappropriate" technology in response to what was perceived to be a persistent failure of the repeated and disparate efforts to implant "modern" technology and techniques into less developed economies. The idea of a technology being inappropriate in a particular setting rests on the idea that, due to multiple structural, institutional and political factors, the actual market prices of inputs in that setting fail to reflect their true scarcity (or shadow) value. Classically, it is argued that the price of (skilled) labour is artificially inflated (e.g. by union power, legislation, restrictive practices), while that of capital (or other technological inputs – e.g. drugs) is artificially lowered (e.g. by foreign donation, tax breaks on capital imports, dual exchange rates etc.). As a result, firms and production units choose a (typically capital-intensive) production technique which is cost-minimising in terms of market prices, but is socially sub-optimal – typically, it will tend to be unsustainable, as it fails to economise on the resource(s) which are genuinely scarce in that setting, under-utilised, and will frequently tend to maintain or exacerbate dualism in the economy (Todaro, 1977; Meier 1995). Considering health care specifically, Kleczowski (1980) argues that an "appropriate" technology is one which is geared to local problems and conditions, i.e.

> "…that technology which is relevant to a given techno-socio-economic framework at a given point of time; it is that technology which contributes the most to the social and economic aspects of development…it should be scientifically sound, acceptable to those who apply it and to those for whom it is used, and affordable to the nation."

The issue of "appropriateness" is raised in order to identify i) the possibility that cost-minimisation may have unforeseen impacts, ii) to consider whether a "cost-effective" technology can ever be an inappropriate technology, and iii) to conclude with a brief discussion of the range of health technologies which may truly be available to a developing country.

As indicated above, cost-minimisation with respect to market prices may have unanticipated or adverse welfare effects if market prices fail to capture the true scarcity values of factors of production. The theoretical background to this issue is discussed at length in the project appraisal and cost-benefit literature (see e.g. Squire and van der Tak, 1975; Irvin, 1978). Most importantly, this literature identifies

the following areas as especially prone to distortion or inappropriate valuation: wages, exchange rates (critical in health care given the heavy reliance on imported drugs and supplies), costs of foreign borrowing, and foreign donations. Without becoming diverted by this discussion, it is especially important to note that, even where comprehensive project appraisal takes all these factors into account, the reality of management and planning is that many critical decisions will be taken on the basis of market prices only – who has time to conduct a specialised technical evaluation when the answer is obvious and the Minister wants it yesterday? It thus seems entirely possible that apparently cost-minimising decisions still have room to result in inappropriate technological choices – ironically, perhaps all the more so as decision-making power becomes more decentralised in many systems.

There is also clearly a link between the concepts of a "cost-effective" technology and an "appropriate" technology. A technology evaluated by a properly executed cost-effectiveness analysis, which fully accounts for the potential distortions described above, and which indicates that the intervention or programme is more cost-effective than alternative programmes currently in use, is affordable in the setting in question, and is feasible given current skills and infrastructure, is by definition an appropriate technology. Clearly, the last sentence contained a large number of qualifiers, any number of which may not fully obtain in practice. There seems to be a particular danger in the application of international or generalised cost-effectiveness estimates for decision-making purposes, which may fail to capture several of the local dimensions necessary for a fully-rounded assessment. Generalisation of cost-effectiveness results will necessarily involve the generalisation of a single technique of production (that which was evaluated); even with the use of factors such as health-care purchasing power parities to "adjust" the estimate to reflect local costs, the possibility of different (but still technically and economically efficient) techniques reflecting local relative prices and resource availability is instantly lost, and the application of a "mono-technic" model (there's only one best practice) is implicitly accepted. There are clearly substantial arguments in favour of generalisation of cost-effectiveness results (especially as approaches to improving generalisability develop) – but the continuing possibility that countries may be saddled with inappropriate technologies must be borne in mind at all times.

Implicit both in the concept of appropriate technology and in the conventional discussion of technology choice and technical efficiency is the assumption that firms or countries have a range of techniques to choose from. Indeed, a conventional approach suggests that firms can choose from all known techniques – from the latest

invention back to the very first process invented in the relevant area – and that past techniques, however antique, will never disappear (e.g. Lancaster, 1969). However, Meier suggests that the menu of techniques actually available to a particular country at a particular time may be very much smaller than the full historical spectrum of "known" techniques. Typical constraints may include a bias amongst foreign suppliers to produce the technology currently in use in developed countries, and to neglect products which best suit the needs of developing countries. This tendency was demonstrated in its most extreme form during the crisis which followed the decision by Aventis to quietly discontinue production of Ornidyl, the first-line treatment for sleeping sickness – despite significant demand for treatment for this very serious disease in several African countries. As an aside, the assumption that viable technologies can never disappear does not seem entirely waterproof. In the mid-1990s, faced with growing problems of multi-drug resistant tuberculosis, the British NHS decided to review possible non-drug treatments, for use either in patients where drug therapy had categorically failed or as a contingency should rates of resistance rise dramatically. Questioned on possible surgical interventions, older thoracic surgeons noted that their teachers had frequently mentioned TB surgery, but no practicing consultants could remember having been taught these techniques. A number of retired thoracic surgeons were identified, who were able to describe and document the techniques they had last used in the early 1950s. Had this review happened ten years later, one might well have been forced to say that this technology was no longer "available". More fundamentally, though, it is important to note that the powerful forces of suppliers' desire for easy profitability, and the tendency of medical culture to aspire to "international best practice" may both push developing countries towards repeated choices of inappropriate technologies. It is therefore particularly important to note that the current debate on the desirability of providing highly active anti-retroviral therapies for AIDS care in African countries (and the many schemes for "free" or discount drugs being bounced around as part of that debate) may contain all the ingredients for an inappropriate and unsustainable technology choice.

Measuring Efficiency and the Potential for Improvement

Measurement Issues

Theoretically, in order to measure the absolute technical and economic inefficiency of a production unit, we would need to know the underlying production and cost functions for that unit. This requirement poses significant problems for real-world application. First, the extreme heterogeneity and complexity of health care interventions (especially at the level of a large, multi-product production unit such as a hospital) effectively rules out the development of engineering-type production functions for all but the simplest interventions. If bottom-up engineering functions cannot be described, then clearly some form of statistically-derived estimation from observed data becomes necessary. However, we can only assume that a statistically estimated production or cost function reflects the underlying, "true" function if we assume that production units are always technically and economically efficient in their operation. From our earlier discussion, we clearly have good reasons to conclude that health care production units are unlikely to meet these conditions in reality. As a result, we must accept that the efficient unit isoquant / isocost line is unobservable, and that any estimated production or cost function cannot be assumed to represent the production frontier or the underlying cost function (McGuire, 1987).

In practice, then, efficiency measurement in health care is almost always going to measure *relative* efficiency – that is, efficiency relative to some benchmark comparator (the best or some sample of best achievers in whatever dimension of efficiency we choose to examine), or changes in efficiency over time. This is an important limitation, which should not be under-stated. Depending on the measurement technique we choose, we can estimate the gains from improving aggregate efficiency up to that of the benchmark unit with a degree of confidence. Once at the benchmark level, we certainly should not assume that no further potential for efficiency gains exist – but their existence and realisation become purely matters for speculation.

Measurement Methods

Adapting the categorisation of Barnum and Kutzin (1993), it is possible to categorise the main measurement approaches relevant to estimating efficiency and the scope for improving technical and economic efficiency. They identify three main approaches: input / ouput ratios and performance indicators; accounting costs; and

statistical cost and production functions. To this list should be added frontier estimation methods.

Input / Output Ratios are, strictly speaking, measures of partial factor productivity. They can comprise any ratio of input to output quantity. Frequently used examples include: visits per nurse / doctor per day; throughput (number of cases) per bed; films taken per X-ray station, and length of stay (total occupied bed days divided by number of admissions). They suffer from the general weakness of partial factor productivity measures that gains or losses in productivity may be imputed to the input in question when, in fact, they result from changes to other inputs. **Performance indicators** typically used in conjunction with such ratios tend to measure either capacity utilisation, or achievement of some form of mission-oriented target. Key indicators of capacity utilisation include measures such as bed occupancy rates and operating theatre utilisation rates; related corollaries of utilisation include turnover interval (the average time a bed is empty), surgical cancellation rates, machine downtime etc. Non-utilisation targets include such indicators as percentage of surgical procedures performed on an ambulatory basis, immunisation coverage rates, cervical screening rates etc. It is absolutely essential to distinguish between such output-related indicators, and purely input-related measures, which are typically expressed in terms of rates per population (e.g. beds per 1000 population, doctors per 1000 population etc.) and from which nothing can be directly inferred regarding efficiency. Certain input ratios (e.g. nurse:doctor ratio) may be indicative of inappropriate resource balances, but cannot be taken as *prima facie* evidence of inefficient input mixes without some investigation of relative prices.

Accounting Cost studies allow the calculation of average costs per output (typically per admission, bed day or out-patient visit for hospitals, per clinic visit for primary care). They thus offer the potential for identifying low cost providers, which can be used as a benchmark against which to judge less efficient providers. Barnum and Kutzin identify two main approaches to the calculation of accounting-based costs – step-down costings of individual institutions, and high-level average costs derived from aggregated data for multiple institutions. Step-down costings attempt to assign detailed costs to quite a low level of service provision (per department or specialty, perhaps even per procedure), by using various methods of allocating direct and overhead costs to a particular end-user department. They tend to be detailed and resource-intensive, thus inherently limiting the number of units examined in any given study. Clearly, the fewer the number of units to compare, the more difficult it will become to make any judgements about relative efficiency. There may be some

scope to calculate "ideal" costs from step-down data (e.g. by assigning target values for occupancy, length of stay etc.), but there can be no real confidence that such estimates are achievable in practice. Aggregate data, by contrast, allows more scope for comparing relative performance in terms of average costs – but loses a significant degree of discrimination relative to step-down methods, in that one can no longer differentiate resource use between different uses. Critically, aggregate hospital cost data requires some assumption of a relationship between the cost of out-patient versus in-patient outputs (e.g. rules of thumb such as the South African convention that one out-patient visit is equivalent to the cost of one in-patient bed day), which are frequently rather shaky. Significantly, both methods may be seriously affected by differences in underlying case-mix and severity between institutions, which might reasonably be expected to influence costs. Many methods exist to adjust for case-mix and severity – but all are only useful in environments where significant patient-based datasets are available, a condition which is rarely met in most developing countries.

Statistical cost (and production) functions attempt to estimate an underlying cost (production) function from cross-sectional or time-series data. While it is accepted that the estimated functions derived from such studies do not represent "efficient" production, this approach has particular value in terms of identifying the behaviour of marginal costs st different output levels, and of drawing conclusions regarding the existence and importance of returns to scale. The general technique of multiple regression analysis which is employed in these types of studies also lends itself to the inclusion of large numbers of independent variables, whose potential impact on cost can thus be estimated. This ability has been particularly prized for allowing the incorporation of possibly complex adjustment for case-mix factors. However, this approach is not without its problems. Crucially, as the true functional form is not known in advance, there is always the inherent risk of mis-specification – and a mis-specified function will yield misleading results. Attempts to use more flexible functional forms (Breyer, 1987) sacrifice capability to adjust for case mix or other independent variables. There is also some debate as to whether such studies yield long-run or short-run functions, and what this may imply for their interpretation (Aletras, 1999). Most important for the consideration of efficiency and efficiency improvement is the criticism that the use of central tendency techniques inherently produces an analysis of average performance – i.e. not even best performance amongst inefficient producers (Rosko and Chilingerian, 1999).

By contrast, **Frontier Estimation Methods** entail the estimation of an efficiency frontier (or envelopment surface) from observed sample data, based upon best performance within the sample. Measurement of the deviation of individual production units from this frontier allows the calculation of relative efficiency scores, and the computation of potential efficiency gains if all units could achieve best performance levels. Zere *et al*, 2000, and Knox Lovell and Schmidt, 1988 provide comprehensive discussions of methodological and theoretical issues. Two main variants of this approach should be mentioned. *Data Envelopment Analysis* (DEA) was the original (and still most widely used) form of frontier estimation, but has the drawback of effectively assuming that all variation from frontier production is due to inefficiency on the part of the production unit concerned – when, in fact, some part of such variation in performance is quite possibly due to factors beyond the control of management, and hence could not be eliminated by improved performance. To attempt to counter this shortcoming, *Stochastic Frontier Estimation* (SFE) methods were developed, which use statistical techniques to exclude the effects of random variation, measurement error and exogenous shocks from the estimated inefficiency score. This approach consistently produces estimates of inefficiency which are significantly smaller than those generated by DEA methods, and hence appears to be an important improvement; however, SFE approaches require larger sample sizes, impose a functional structure, and have more difficulty handling multiple outputs. Neither SFE and DEA approaches are immune to problems of model specification; a sensitivity analysis (Ozcan, 1992) indicated that different specifications of a DEA model could yield significant variation in mean efficiency scores (from 67.2% to 91.8%) for the same sample of hospitals.

Benchmarking and Performance Management

The preceding discussion has focused on techniques of measurement. Clearly, though, by itself measurement achieves nothing beyond the generation of interesting (and, frequently, not so interesting) statistics. It is therefore very important to place efficiency and productivity measurement firmly in the context of an active framework of improvement and innovation, and not to see them as static, academic pursuits. This requires some mention of the concepts of *benchmarking* and *performance management*. The concept of benchmarking has grown in significance in the last two to three decades. It is rooted in the very simple idea that organizations can seek comparative advantage (or to remedy apparent disadvantage) by looking outside themselves and learning from the best practice of others (Holloway *et al*, 1995). Having originated in larger private sector corporations, benchmarking has

subsequently been widely adopted by organisations (and regulators) operating in non-competitive environments (e.g. monopoly utilities, and certain developed country health systems, especially the UK National Health Service (Grout *et al*, 2000). Holloway *et al* distinguish between *benchmarks*, which are best practice standards towards which organisations should move, and *benchmarking* as a process, which requires organisational commitment over time to identify benchmarks, then to develop and implement techniques and plans for achieving those benchmarks on an ongoing basis. They suggest that benchmarks may be internal (e.g. comparing departments which share some performance parameters); functional – comparing similar business function across different organisations or industries (e.g. performance of transport and logistics, human resources management etc.); competitive – comparing performance with direct competitors; or generic – comparing performance with "best in class" firms in various industries for particular aspects of operations. Most importantly, they state that an optimal benchmark must be readily measurable and quantifiable; it must be meaningful, relevant and worthwhile; as simple as possible; and be potentially available in external or competitive environments.

In a number of public sector environments, there has been a tendency to move beyond the essentially voluntary concept of benchmarking described above, and to move to an active, hierarchical system of *performance management*. Such arrangements typically involve a higher management level selecting productivity and efficiency benchmarks – often in parallel with quality and outcome benchmarks – which are then translated into explicit targets for organisations and individual managers. More or less "high-powered" incentives may be directly attached to these targets (Grout *et al*), involving direct financial reward for the organisation or individual concerned – and/or "name and shame" approaches, such as publication of league tables, which aim to motivate better performance through personal and corporate desire to avoid opprobrium and perceived failure. There has been considerable experiment in the UK National Health Service involving direct productivity and cost targets (prompted largely by the almost immediate failure of the "internal market" of the early 1990s to have any impact on costs and efficiency), such as the Labour Productivity Index, the NHS Efficiency Index, and the use of "reference costs" to judge average specialty and procedure costs. Many problems have been identified with the specific techniques applied in the NHS (see e.g. Appleby, 1996), mainly centring on its particular proclivity for complicated multi-indicator indices. Grout *et al* discuss in detail possible responses to performance management regimes, and

identify two sets of potential unintended consequences which may result (whilst accepting that well-designed system might equally well achieve its desired outcome of improved efficiency). Managers may tend to concentrate on key (measured) areas of perfomance and neglect others; or (and) they may attempt to "manipulate the signal of performance", through fiddling the data or through gaming behaviour.

Notwithstanding of the potential drawbacks described above, some form of performance management is, of course, a *sine qua non* of improved efficiency in a non-market system that contains no inherent rewards for raising productivity and reducing costs. Cynics might also suggest that the complete absence of effective performance management is a defining feature of most developing country health systems. This important dimension of efficiency improvement will be returned to later in the paper.

The Scale of Inefficiency in Health Care Production – International Evidence

General Approach

In order to gain a more tangible sense of the likely scale of technical and economic inefficiency in health care production, and to allow some suggestion of the likely potential for savings from improved efficiency, the international published literature (along with accessible grey literature) was reviewed. The results of this review are summarised below, and are organised as follows. First, international findings relating to the efficient *utilisation* of services and facilities are summarised, with some discussion of the implication of these results. Second, the much more limited literature which explicitly discusses factor substitution is described. Third, the key literature containing results on the existence of economies of scale and scope in health care production is summarised. Finally, the results of all the studies located which provide explicit estimates of potential savings from improved efficiency are summarised (in tabular form), allowing consideration of what assumptions of inefficiency and efficiency improvement might reasonably be generalisable.

Inefficient Hospital Utilisation

In all aspects of efficiency measurement, hospitals tend to be the best studied component of health systems internationally – not unreasonably, as they account for a majority of health spending in all countries. Barnum and Kutzin (Table 3.3., p94-5) provide a useful summary of hospital utilisation statistics from selected developing countries from which data were available (all during the 1980s). Above all, their data

demonstrate the significant variation in utilisation and practice between countries. For example, their summary results (themselves means, thus hiding even greater intra-country variation) showed a wide range of bed occupancy rates from a low of 46% (Fiji and Turkey) to 125-129% (i.e. overcrowded) in Lesotho, and mean lengths of stay ranging from 3 days to 22 days. It is, however, important to note that it has proved persistently and consistently difficult to obtain comparative utilisation statistics for developing countries on a routine, ongoing basis (Peters *et al*, 2000). More recent comparison of trends in OECD, former Soviet and Central and Eastern European countries (Hensher *et al*, 1999) confirm the existence of substantial inter-country variation, but also identify a clear downward trend in length of stay amongst all three groups of countries over the period 1986-1995. Reductions in length of stay have been particularly important features of the health care environment in both the USA, where average length of stay fell 22% between 1989 and 1998 (Ashby *et al*, 2000), and the United Kingdom, where average length of stay fell 45% for the between 1986 and 1995 (Hensher, 1999a). The achievement of these reductions in length of stay was predicated on the discovery that a very significant proportion of hospital bed days (and, to a lesser extent, admissions) were not clinically necessary. Thus, studies in the US, UK, Italy, Spain, Canada, Ireland and Australia have all identified significant proportions of "inappropriate" bed days, largely relating to the tail end of hospital admissions. This evidence is summarised and reviewed in Edwards *et al* (1998) and Hensher *et al* (1999b). Using a range of tools, these types of studies typically identify from between 10% to as many as 50% of bed days as being inappropriate, and hence potentially being dischargeable or transferable from hospital. At the same time, ability to reduce length of stay has been greatly enhanced by the rapid development of ambulatory or day case surgery and investigation (for example, a 250% increase in day case activity in the UK NHS between 1982 and 1995 – Hensher and Ewards, 1999), which has allowed patients who would previously have required hospitalisation to receive treatment or diagnosis and return home on the same day.

Reducing length of stay and improving bed occupancy can undoubtedly have a significant impact on resource use. Indeed, we can be fairly confident in making a clear and categorical statement about bed occupancy. In acute hospitals, an optimal average bed occupancy rate lies in the region of 85%; rates much below 80% are clearly inefficient, while average rates over 90% give rise to an increasing probability that, on any given day, the hospital in question may have insufficient beds available to meet random daily fluctuations in demand for care. However, the waters are much

muddier with regard to length of stay. Generally, there has been little attempt to apply the techniques of utilisation review and appropriateness evaluation in developing countries – so we cannot be sure that the same "tail" of *inappropriately* long hospital stays exists in developing countries. Second, available resources and technology may legitimately impact upon lengths of stay. Thus, patients living in deep rural areas with no access to primary health care may be retained in hospital for observation longer than they would if they lived in an urban area with good primary care – because, for example, a post-operative complication that could rapidly be picked up and treated by a primary care clinic can become a life-threatening event where no clinic exists. Yet they may just as well be staying in hospital longer because no patient transport is available to take them home. Equally, patients in poor countries may be receiving older vintage treatments, which require longer hospital stays. Outright resource shortages can lead to longer stays – shortages of key essential drugs mean that severely sick people may be receiving little more than palliative care – in which case they will stay in hospital until they either get better or die, whereas basic drug availability could have ensured they were on the road to recovery and out of hospital much sooner. We should therefore be rather more cautious in our efforts to establish international benchmarks for length of stay and other related indicators. We are, however, on rather stronger ground where we are able to generate performance comparisons and benchmarks from local data. Thus, while it might be quite unreasonable to require a Mozambican hospital to operate at American or British lengths of stay, we might consider it much more reasonable to require it to move towards those lengths of stay observed at "best practice" benchmark hospitals of a similar type and case-mix elsewhere in Mozambique. Box 1 provides an example of the use of simple hospital indicators and benchmarking, based upon data from South African district hospitals, indicating what can be done simply through the use of local performance data, with no reference to international benchmarks or comparators.

Box 1 – Performance Indicators for South African District Hospitals, 1999/00

The complex and tortuous process of integrating the thirteen separate Apartheid-era administrative systems into a single South African national health system has made performance data a scarce commodity in recent years. However, data are increasingly flowing successfully, and a near-complete dataset of activity and utilisation statistics for district, regional and tertiary hospitals has, for the first time, been collated for the financial year 1999/00. In this dataset, information was available for 234 out of 264 district hospitals. After excluding 21 very small hospitals (<30 beds) and 20 outlier hospitals for which data were suspect, it is possible to use the cleaned dataset of 193 hospitals to illustrate the potential impacts of basic efficiency improvements, all based upon performance levels currently being achieved in South African district hospitals.

Current performance in terms of length of stay and bed occupancy can be summarised for the sample of 193 hospitals with 34,673 usable beds:

Length of Stay Bed Occupancy	
Mean	
	5.7 60.5%
Median	
	4.7 60%
10th Centile	
	3 84%

Clearly, we could set the mean, median and 10th centile (i.e. top 10%) performances as benchmarks for poorer performing hospitals to attain. If we assumed that all hospitals currently below these benchmark levels were able to achieve them, while all hospitals above the benchmarks continued to function at their current levels of performance, we can calculate the reduction in total numbers of beds which would be made possible (with no reduction in output in terms of numbers of admissions).

Potential Reduction in Bed Numbers: LOS Only LOS and Occupancy	
Achieve Mean	
	3,666 6,061
Achieve Median	
	5,440 9,067
Achieve 10th Centile	
	10,089 12,011

The table shows clearly that significant reductions in total bed numbers could be made – indeed, 10% of beds could close simply by the relatively unchallenging target of all hospitals achieving at least the current mean value of length of stay – while achieving the level of performance currently set by the best 10% of hospitals would allow more than one third of beds to be eliminated.

Data Source: Department of Health - Health Financing & Economics Directorate / Pamela Ntutela / Andy Burn

It should also be noted that various sources identify a common theme with regard to bed occupancy in developing country hospitals – namely that bed occupancy is highest in tertiary hospitals and steadily declines as one passes down the descending "levels" of hospital (Barnum and Kutzin; Somanathan *et al*; South African Department of Health)

Primary Health Care Utilisation Rates

While hospitals lend themselves to analysis using utilisation rates and performance indicators, primary health care has proved to be less amenable to utilisation measures. It is far harder to conceptualise "capacity" in terms of clinics and clinic visits, and overall data availability is far poorer. It clearly would be possible, given the appropriate data, to construct productivity indicators for basic primary care, such as consultations per nurse or doctor per day (or other time period), births per midwife etc. – and indicators of this sort are, no doubt, available in certain countries. However, during the admittedly non-systematic and non-exhaustive literature review undertaken for this exercise, no studies were identified which directly used such indicators to compare PHC productivity. Some studies have attempted to consider the proportion of staff time spent on discrete activities (e.g. patient contact, administration, training etc.). Gilson (1995) found that, amongst staff working in primary care dispensaries (both government and church run) in Tanzania, less than 50% of staff time could actively be assigned to identifiable activities. Similarly, Daviaud (1999) found that time spent by primary clinic nurses on direct patient contact varied from 1.9 hours per day to 6 hours per day (in an 8 hour working day) across clinics in the South Peninsula municipality in South Africa – implying that some frontline professional nurses (none of whom were employed as managerial or administrative staff) were spending less than 25% of the working day actually providing health care.

Input Mix and Substitution

The potential significance of opportunities to improve the economic efficiency of health systems through the reprofiling of health workers and/or achieving a more optimal mix between labour and capital has already been alluded to in this paper. Inappropriate balances of human resources and skills are frequently raised as a major source of economic inefficiency (see e.g. Barnum and Kutzin; Hensher, 1998; WHO, 2000), and managers from different systems and countries around the world would undoubtedly concur with the diagnosis and its implied remedy – skill mix reprofiling. Unfortunately, direct evaluations of the scale of such skill-mix inefficiencies are, on closer investigation, remarkably difficult to find. Martineau and Martinez (1996) suggest that, beyond a small number of local studies analysing skill mixes and skill requirements in individual developing countries, we actually have little or no information on the cost of inefficient skill mixes, the cost of shifting towards new staffing patterns, or on what would be the most efficient approaches to retraining staff. Amongst the literature reviewed here, only one study directly tackled this issue. Somanathan *et al* directly compare the ratio of the marginal products of doctors and nurses and the ratio of their wages in different level hospitals in Sri Lanka. Their analysis indicated significant potential gains in output (or cost savings) through reducing the ratio of doctors to nurses.

Still more elusive is any empirical analysis of the potential efficiency gains from optimal substitution between labour and other input classes. Mills (1990b) summarises data on expenditure by category of input for a number of developing countries; her data reveal significant variation in the share of key factors (especially personnel) in hospital spending, and between capital and recurrent expenditure. Clearly, though, these data are purely descriptive means. Similarly, in a very limited way, the 2000 World Health Report (WHO, 2000) attempts to give some impression of relative input mixes. To provide directly useful approaches, however, considerably more thought would need to be given to developing a viable methodology for estimating optimal input mixes, which would only be applicable at country level and which would most likely require rather more discriminating data on activity / quality variables than are typically available at present.

Variation in Average Costs

Clearly, a significant literature exists presenting cost studies and average cost estimates for different levels and types of health care provider in many different developing countries. Barnum and Kutzin (1993) and Mills (1990) amongst others

present comparative data from various studies for unit costs of hospital care, converted into dollar terms. What is less clear, however, is the extent to which cross-country comparison of unit costs can actually provide meaningful information on inefficiency and the scope for efficiency improvements. As already discussed, the outputs of these studies reflect different absolute local factor prices, differing relative prices, and differing input mixes. In the absence of concerted efforts to produce a really convincing health sector purchasing power parity index, the value and validity of direct international comparison of unit costs would seem to be extremely limited. The emphasis thus clearly moves towards using local and country-specific studies to illustrate potential efficiency improvements at local level. It is not proposed that this paper should attempt to catalogue and review all the costing studies which might be relevant – not only because of practical constraints, but also because such studies frequently do not pose comparison of relative efficiency of production units as their primary focus. This is particularly true where costings are undertaken to feed cost-effectiveness analyses; such studies tend to focus on providing a point estimate of the cost of a given service or intervention, when wider consideration of efficiency requires comparison of a sample of several production units. Various of the studies already identified provide some insight into the potential value of local cost comparisons. Key aspects include the general and consistent pattern of significant variation in unit costs between production units when larger samples are considered. Thus, for example, Somanathan *et al* encountered eight-fold variation in costs per admission to their sample of fifteen "complex" Sri Lankan hospitals, six-fold variation in cost per admission to thirteen "intermediate" level hospitals, and a very similar scale of variation in costs per patient day. The South African district hospitals dataset, even once trimmed for outlier values, still returns a near-fivefold variation in costs per patient day. Clearly, then, where it is available, unit cost data may present a powerful basis for benchmarking and for identifying relatively inefficient units. It is also important to note that step-down costing methods offer some potential for analysing sources of inefficiency and for modelling potential improvements – a property which is not shared by larger aggregate datasets (such as the South African data), which provide data only on total expenditure. This leads directly to a major potential constraint upon the feasibility of using accounting cost data for efficiency improvement programmes – namely whether the data required can be produced systematically and repeatedly on an ongoing basis, or whether they are only practically available on a one-off basis, requiring significant effort to assemble a single cross-sectional dataset. It is certainly possible in developing countries to develop and maintain aggregate data on expenditure at provider unit level (c.f. the

South African data), which can be produced periodically to illustrate trends over time. The major confounding factor in such datasets tends to be the definition of types of provider units, and the inadvertent aggregation of data on lower level units which are under the managerial control of another. Thus, for example, a district hospital may be responsible for the management of primary care clinics in its district, and their expenditures may in fact be included in the aggregate expenditure reported by the hospital. By contrast, it is extremely difficult to replicate detailed step-down cost analyses on a routine basis – but without comparisons over time, an individual unit clearly cannot track its own performance improvements. To do so requires the development of cost-accounting and cost centre management techniques at the level of the production unit. These do not need to be particularly sophisticated – but the trick comes in defining, applying and maintaining uniform accounting definitions and rules to allow comparison of costs across different providers. It is this latter dimension which has proved to be more difficult to implement in developing (and, indeed, developed) country settings.

Economies of Scale and Scope

Over the last ten years, a relatively clear and unambiguous consensus has emerged in the international literature on the existence of returns to scale in the hospital sector. Most critically, a large-scale systematic review of the issue (CRD, 1996), which examined over 100 studies, concluded that cost-function studies yielded a consistent message:

> "Almost without exception, those studies which use cost per case as the unit of analysis, and in which case-mix adjustment is adequate, show evidence of constant cost of diseconomies of scale." (CRD, 1996)

Further, DEA analyses also yielded highly consistent results, namely that hospitals with fewer than 200 and more than 620 beds are scale inefficient. Commenting on these findings, Posnett (forthcoming) articulates these findings into a direct policy recommendation – any economies of scale that may exist in hospital production are exhausted at a low level (100 to 200 beds maximum), and diseconomies of scale seem to be a significant feature of hospitals much over 600 beds in size, suggesting that the "optimal" hospital size is probably in the region of 300 to 600 beds. It should be noted that the same author (Posnett, 1999), discussing the same results, gives a slightly different suggested optimal size, namely 200 to 400 beds (however, this difference of interpretation does not change the basic drive of these results). This

view seems to be confirmed by more recent developing country studies, such as that of Zere *et al* (forthcoming), which found that, amongst South African hospitals in the former Cape Provincial Administration, there was evidence of increasing returns to scale amongst smaller level I district hospitals, and of decreasing returns to scale amongst (larger) level II and III (regional and tertiary) hospitals.

The CRD review also concluded that the evidence regarding the relationship between clinical outcome and patient volumes (which has at times been cited as a rationale for concentrating services at larger hospitals) is weaker than often believed (due in large part to major methodological defects in key studies), that where such a relationship can be demonstrated, it often implies a relatively low minimum threshold volume (with no further improvement from increasing volume above this threshold), and that, in any case, any such relationships relate only to the size of a particular department – and say nothing about optimal hospital size. It found also that evidence of the existence of economies of scope was relatively limited, and of little direct policy significance – echoing Barnum and Kutzin's assertion that economies of scope "…are not an important factor in planning hospital services."

No studies were located in the present review that might have yielded generalisable results on economies of scale of relevance to primary health care or non-hospital services in developing countries. However, a British study (Giuffrida *et al*, 2000) of FHSA (primary care management agencies) efficiency and a US study of Area Agencies on Aging (planning and management of services for older people - Ozcan and Cotter, 1994) both indicated the presence of economies of scale (in terms of size of population covered) for health management agencies.

Potential Savings From Reduced Inefficiency – Direct Estimates

A number of studies were identified which generate direct estimates of the scale of inefficiency observed in a range of health care production environments. These mainly employed DEA or SFE methods (which have the great advantage of producing efficiency scores, which are directly equivalent to the potential increase in output or cost saving achievable were the average unit to attain the level of efficiency seen in benchmark "frontier" units). A few are from developing and middle-income countries, but most are from developed countries. It was felt useful to summarise the results of both groups, in order to give a sense of the range and scale of inefficiency observed across an extremely diverse set of environments. Table 2 summarises the results for developing and middle income countries, and Table 3 those for developed countries.

It is extremely important to note that the sample of studies presented below do not represent the findings of a systematic or exhaustive literature search – but they do represent a broad spectrum of this type of work, identified via a reasonably thorough search. Clearly, caution must be exercised in their interpretation; most critically, there can be no guarantee that the estimated levels of inefficiency, if correct, can necessarily be eliminated. Equally, all of these studies measure relative inefficiency – distance from the current production frontier as derived from actual practice – and therefore can say nothing about what future improvements in efficiency could be achieved by shifting out the frontier (and none would claim that the relatively "efficient" providers in their samples could not themselves improve their efficiency of production). Their function here is to serve as a guide and illustration of the general magnitude of technical inefficiency encountered when one goes looking for it armed with the appropriate methodological tools.

Table 2 – Inefficiency Estimates from Developing and Middle Income Countries

Country	Subject	Method	Savings / Output Increase Possible	Reference
Sri Lanka	Public Hospitals	Production function Cost function	Level of Hospital: Complex – 44% output increase Intermediate – 25% " " Basic – 100% " " Complex – 26% saving Intermediate – 21% saving Basic – 24% saving	Somanathan *et al*, 2000
South Africa	Public Hospitals	DEA	Level of hospitals: Level I – 26% cost saving and 30% bed reduction Level II & III – 33% cost saving and 39% bed reduction	Zere *et al*, Unpublished
Turkey	Public Hospitals	DEA	9.4% cost saving *or* output increase	Ersoy *et al*, 1997
Kyrgyzstan	Public Hospitals	Costing and econometric model	Reduced LOS and bed reductions without hospital closures – 19% cost savings Reduced LOS and hospital closures – 43% cost savings	Street and Haycock, 1999
Tanzania	Public dispensaries	Costing	All units achieve median performance: 8% cost saving on personnel 4.5% cost saving on drugs	Gilson, 1995
Sub-Saharan Africa	Drug use and procurement	Model	Implementing range of efficient procurement, distribution and prescribing practices: 10 – 60% cost savings	Foster, 1991

DEA – Data Envelopment Analysis; SFE – Stochastic Frontier Estimation

Table 3 – Inefficiency Estimates from Developed Countries

Country	Subject	Method	Savings / Output Increase Possible	Reference
Denmark	Public Hospitals	DEA Free Disposal Hull	41% input saving / output increase 2% " "	Holvad and Hougaard, 1993
Sweden	Public Hospitals	SFE	Switch to output-based reimbursement: 9.7% cost saving	Gerdtham *et al*, 1999
USA	Pennsylvania acute care hospitals	SFE	Without case-mix adjustment: 18% cost saving / output increase With case-mix adjustment: 8% cost saving / output increase	Rosko and Chilingerian, 1999
USA	Michigan Public and Not-for-Profit Hospitals	DEA	Public hospitals: 2.2% input saving / output increase Private Not-for-Profit hospitals: 11.9% input saving / output increase	Valdmanis, 1990
USA	California Catholic Hospitals and non-Catholic not-for-profit hospitals	DEA	Catholic Hospitals: 19% input saving / output increase Non-Catholic Hospitals: 24% input saving / output increase	White and Ozcan, 1996
USA	Nursing Homes	DEA	7% input saving / output increase	Nyman *et al*, 1990
USA	Physician practice in treating *otitis media*	DEA	Comply with optimal treatment protocol 20% input saving / output increase	Ozcan, 1998
USA	HMOs	DEA	33% input saving / output increase	Rosenman *et al*, Unpublished
Norway	Public Dental Clinics	SFE DEA	7% input saving / output increase 26% " "	Grytten and Rongen, 2000
USA	Community Mental Health Centers	DEA	CMHCs with inpatient services: 19% cost saving / output increase CMHCs without inpatient services: 53% cost saving / output increase	Tyler *et al*, 1995

Implications of the Research Evidence

This brief review of the available evidence on the magnitude of technical and economic inefficiency in health care systems can be synthesised into a fairly succinct set of critical findings and implications:

- There seem to be very strong grounds for assuming that all health care systems, in both developing and developed countries, will display significant intra-system variations in performance, technical and economic efficiency
- Similarly, there seems to be scope for significant real savings in all or most systems from reductions in relative inefficiency achieved by pulling poor performers up to benchmark performance levels (notwithstanding any scope to further improve the efficiency of "frontier" production units)
- That, with certain exceptions (e.g. bed occupancy rates), international and inter-country benchmarks and comparisons for technical and economic are actually limited in value – in particular, inter-country comparisons of unit costs are unlikely to yield meaningful conclusions
- Our ability to estimate the impact of inefficiencies relating to sub-optimal input mixes is currently very limited – even though this in no way diminishes the likely importance of this issue
- We know a lot more about inefficiency in hospitals (and how to tackle it) than we do about inefficiency at other levels of care
- There are strong grounds for accepting that hospitals which have less than 200 beds are too small in terms of scale economies, while hospitals which have over 600 beds seem very likely to display diseconomies of scale

The direct implications of these conclusions are diverse, but also need to be spelled out. The scale of technical and economic inefficiencies from the developing country studies displayed in Table 2 is worth contemplating; most of the hospital studies indicate potential savings of 20% or more relative to current resource use. Given that the share of hospital expenditure in total health expenditure averaged 30-50% in the developing countries surveyed by Mills (1990a), and 30 to 81% of total public health spending in those surveyed by Barnum and Kutzin, hospital inefficiency could easily tie up the equivalent of 10% or more of total health spending – a significant prize, even if only half that level of savings could be unlocked. Given that we also have a far less developed understanding of the scale and nature of inefficiency in primary

care, it therefore seems reasonable to suggest that the hospital sector (at all levels) should be our main focus, certainly in terms of immediate action.

Still on the theme of hospitals, the evidence on optimal hospital scale raises two major concerns. First, it calls for a rethinking of the formulation of the district hospital concept. District hospitals are frequently small; indeed, small hospitals with less than 100 beds are still being developed in certain countries (e.g. Tanzania; Flessa, 1998). There is no denying that many situations will arise, especially in remote communities with poor transport links, where it is entirely unavoidable and quite desirable for a "sub-optimally" small hospital to operate. Equally, though, there are likely to be many situations in which hospitals are too small, operate inefficiently, and offer a poor scope of service – and where health care needs might well be better served by a process of consolidation and simultaneous improvement of patient transport services. As a minimum, development of future district hospitals should start with the presumption that they should be larger than 200 beds, with the burden of proof lying with those who wish to have more, smaller hospitals. Second, the strong indication of diseconomies of scale at larger sizes raises questions about the future of a number of African dinosaurs – for example, University Teaching Hospital, Lusaka, and Chris Hani Baragwanath Hospital, Johannesburg, which both weigh in over the 1300 bed mark. As an absolute minimum, future developments (particularly of "prestige" tertiary hospitals) should be very deliberately constrained to the suggested 600 bed maximum – and that injunction must surely include any planned replacements for the existing very large hospitals.

More generally, the evidence suggests that we should be very mindful of the potential impact of demand upon utilisation – especially where health systems and current organisational models are clearly failing to give users what they want, leading patients to vote with their feet and health facilities to operate at inefficiently low levels of capacity utilisation. This theme recurs repeatedly – e.g. poor drug availability and recurrent shortages in government clinics mean that patients do not bother to present (Gilson, 1995; Bitran, 1995); lack of skilled staff, especially doctors and specialists, leads patients to bypass lower-level hospitals, resulting in very low utilisation rates (Barnum and Kutzin, 1993). A compromise is required here – resources must be found to improve the availability of key inputs (drugs, supplies, physicians) at some sites, while, potentially, other sites may have to give up the pretence of offering that service (e.g. downgrading small, understaffed district hospitals to become health centres). It is precisely at this point that managerial capability – the ability to find

successful ways of getting difficult things done – starts to be far more important than technical analysis and evaluation.

The clear implication that the role of international benchmarking is limited, and that the core issue is relative inefficiency within individual systems, has its own implications. In the project of improving health system technical and economic efficiency in developing countries, there can be no substitute for local data collection, analysis and action. No amount of international collaborations and comparative databases will get around the need for sleeves up, hands-dirty field work; and sophisticated methodological development and debate can probably have only the slightest of impacts upon the much more complex challenge of establishing and maintaining simple, robust and meaningful data systems, which can generate the information required to make basic performance management a possibility. There will be ample room for international expertise and assistance – but it must surely be in the form of training, capacity building, and learning by doing. Perhaps the one exception to this homily lies in the need to develop more adequate and operationally viable methods for assessing optimality and sub-optimality in overall input mix and personnel skill mix. The review seems to indicate that this area is ripe for development – crucially, in developing analytical tools that would allow comparison of different input combinations and prices and illustrate the relative efficiency of alternative options (and, ideally, to suggest alternatives to current practice which otherwise might not be identified without external prompting).

A Proposed Framework for Addressing Technical and Economic Inefficiency

The preceding discussion of the nature of technical and economic inefficiency in developing country health systems, the identification of these phenomena as essentially relative concepts, and the glimpse afforded by the current literature of the likely scale of the problem all suggest certain approaches to tackling the question of inefficiency in the production of health care. Most importantly, the relative nature of the concept requires that each country must develop a strategy of its own – and that, in turn, each health care provider unit or facility develop its own efficiency improvement programme. There is much room for sharing of experience and expertise – both in measurement and in implementing efficiency improvement measures – but there is fundamentally no way out of the need to identify specific problems from the top to the very lowest level of the system, and to develop solutions which will fit local realities and overcome very particular local obstacles. It is

proposed that a successful national-level approach to developing an efficiency improvement programme would contain the following components (which are not all sequential steps):

- Identification and quantification of major areas of technical and economic inefficiency and potential gains from efficiency improvement
- Assessment of priority employment of funds / resources released through efficiency improvements
- Identification of key causes of identified inefficiencies
- Assessment of possible interventions to improve efficiency
- Assessment of likely constraints acting upon efficiency improvement options, and estimation of likely scale of savings realisable
- Implement structural changes required to facilitate major or one-off improvements
- Implement organisational and cultural shift to continuous productivity improvement, including appropriate performance management and incentive systems

Identifying and Quantifying Major Inefficiencies

The review of methods and studies provided in this paper gives a clear sense of the range of techniques available for deployment in the search for inefficiencies. Most critical, however, is the development, full implementation and subsequent maintenance of a basic data reporting system, which provides useful, meaningful information on activity, expenditure, productivity and efficiency. This boils down to a robust hospital data system which will provide, *as a minimum*, aggregate numbers of admissions (or deaths and discharges), bed days (and hence mean length of stay), available beds (hence bed occupancy), total surgical procedures, total radiological procedures, day cases, outpatient visits and total expenditure per hospital, according to a uniform and well-understood set of definitions. Possible "extras" might include data on staffing full-time equivalents by functional categories, and/or sub-division of data by broad specialties. Case-mix data should not be treated as a priority unless a viable system is already operating successfully. For primary care, the data reporting system should concentrate upon total attendances per facility, with some programmatic breakdown (e.g. antenatal care, postnatal, immunisation, children general, adult general, family planning, STD etc.), staffing full-time equivalents by

main functional categories, and expenditure per facility. Sounds familiar? These data, or something very like them either already are or theoretically should be collected in most health systems. The critical first step, though, is to systematise the data collection and transmission process, so that the data really are provided and collated, with a basic level of confidence in their quality and comparability.

Once these basic data become available, clearly the array of techniques for efficiency measurement presented earlier can all be deployed. Without regularly available routine data, however, we are forced to rely on one-off sample data and special studies. Stop-gap approaches to data militate strongly against measurement of progress and improvement over time, and generally fail to cover all providers – both serious impediments to the process of efficiency improvement. It is almost certainly preferable to obtain maximum coverage of even a very crude data system than it is to focus on obtaining more sophisticated data at pilot sites – because without the former, no analysis will be possible at any sites other than these pilots.

Excluding the more analytically demanding approaches (e.g. cost function estimation, DEA), which are likely to require very specialised technical and academic expertise to employ, the most critical issue becomes choosing an appropriate set of benchmarks. Considerable thought needs to be given to this issue, possibly with a particular focus on choosing benchmark providers which have good all-round performance (i.e. perhaps not the best in terms of any single variable, but in the top 10% for all etc.). There is room here for methodological development work, perhaps using the more sophisticated approaches to investigate which basic indicators are the most consistent and reliable predictors of inefficiency.

Assessment of Priorities for Additional Resources

The earlier discussion of interactions between technical, economic and allocative efficiency identified the importance of having some shared understanding of what the priorities will be for any resources released through efficiency improvements. Where more sophisticated sectoral resource allocation processes are being developed (e.g. application of sectoral cost-effectiveness analysis) this is likely to be quite straightforward, in the sense that analysis already undertaken can be used directly. In the absence of more sophisticated work, some discussion will need to take place regarding the stated health priorities of the country, and their likely fit with the sorts of resources which are likely to become available given the nature of the inefficiencies that have been identified. The core question here is to ask whether we want more of the same (i.e. increased output for current inputs), or to release resources for other

uses (current output for reduced inputs), in order to plan our efficiency improvement measures accordingly.

Identification of Causes of Major Inefficiencies

It is obviously essential to understand why particular inefficiencies are arising if there is to be any realistic chance of reducing them. Application of the constraints / incentives matrix presented in Table 1 can offer a means of conceptualising likely contributors to inefficiency. It should probably be accepted that this is a qualitative exercise. Most critically, though, the people responsible for the inefficient services (be they clinical or administrative staff) are likely to be the best source of insight into causes of inefficiency. Whether more formal qualitative research methods are used to elicit their views, or whether managers simply spend time to ask questions and listen to opinions, those who are caught up in the heart of inefficient practices must be questioned in detail about why things happen as they do and, critically, how things might be improved. A significant portion of technical inefficiency (i.e. wasteful processes) relates to extremely micro-level clinical and administrative custom and practice, which generalist managers or researchers may not necessarily be able to identify as inefficient. Overdyk *et al* (1998) provide a fascinating discussion of the extremely micro-level changes in scheduling, organisation and day-to-day operation which they undertook to achieve significant improvements in the efficiency of their operating rooms, involving a level of intervention that no centralised policy could effectively capture.

Identifying potential remedies to inefficiencies requires a two-track process. At one level is the grass-roots approach of involving workers and stimulating process improvements and initiatives by all those involved in the process of care delivery. However, there is, of course, an extensive stock of experience and knowledge already available internationally, which can be drawn upon to provide rather more fundamental changes and innovations. For example, a great deal is now known about practical measures which can be taken to reduce length of stay and to avoid inappropriate admission; some attempts have already been made to summarise this knowledge (see e.g. Edwards *et al* 1998). A greater effort to disseminate and diffuse this operational knowledge base, perhaps in a straightforward “how to” style, and certainly using the internet as a key mechanism, would probably be of significant assistance to managers on the ground.

Assessment of Possible Interventions and Likely Savings

Potential interventions to address inefficiencies must then be identified and assessed in terms of feasibility and potential pay-off. Clearly, there may be choices between simpler strategies which may not release the maximum conceivable saving, and complex interventions which do – but with the key question being real-life feasibility and risk of failure (better to have an boring plan and some savings than a brilliant plan and no savings at all). Feasibility of implementation is likely to be affected by a range of constraining and enabling factors. Table 4 attempts to summarise some of the key influences upon the feasibility of efficiency improvement initiatives.

The potential importance of attitudes to job losses cannot be over-stated. Where major inefficiencies have been identified (e.g. on the scale pointed to by the developing country studies in Table 2), it is highly unlikely that equivalent savings can be realised without job losses. WHO (2000) notes that "tensions" may arise between managers and politicians when the right to shed workers is withheld due to political pressure. It is particularly important that politicians understand that they will not be able to have both savings and no job losses, and that squeezing non-personnel funds is likely simply to exacerbate existing inefficiencies.

Table 4: Factors Likely to Impact on Feasibility of Efficiency Improvements

Constraints	Enabling Factors
Job losses politically unacceptable	High-level political commitment to "do what it takes" to reshape system
Lack of political will to change or delete restrictive regulations / legislation	
Legal framework of employment precludes compulsory redundancies of workers	Need for efficiency improvement seized as an opportunity to update procedures
Inflexible budgetary and/or procurement systems	Maximum budgetary flexibility and delegation is permitted
Professional associations have not bought in to need for efficiency improvements	Political capital invested to bring key professional groups on side
Corruption widespread and reforms do not include increased salaries to counterbalance loss of illicit earnings	Clear rewards for abandoning previous corrupt practices, and genuine commitment to eliminating corruption at all levels
Poor general infrastructure, especially poor communications, which is not improved to support efficiency improvement programme	Integration of responses and support from other sectors, again requiring political commitment and coordination
Unwillingness to shift resources from other parts of system to leverage efficiency savings	Acceptance that significant savings are unlikely to be realised without some up-front investment (spend to save)
Refusal to commit interim additional funds to leverage efficiency savings	
Perceived risk that savings will be "taken away"	Clear guarantee that savings will be retained
Financial distress (e.g. persistent failure to pay salaries, suppliers etc.)	
Inadequate managerial capacity and lack of commitment to efficiency improvement, trying to reduce "unproductive" management	Strong cadre of high status non-medical professional managers for whom efficiency improvement is a key task

It is also very important to understand that many savings will take some considerable time to realise, and may well require up-front investment. Thus, the training of a new cadre of multi-purpose health workers requires investment in curriculum development, training and recruitment before any of the downstream benefits from more efficient skill-mix can be captured; while shedding excess staff will require funding for redundancy packages, retraining measures etc.

Thus, there may well be ambiguous and difficult choices to be made, which should not be hidden. In addition to politicians' well-known and entirely rational aversion to

job losses, other tensions may well emerge. Importantly, an ideal system incentivises efficiency by allowing providers to retain any savings they secure – but the very objective of using efficiency savings as a means of releasing additional resources may explicitly require that these savings be diverted to different providers and higher priority activities, hence undermining incentives to economise. Not to be underestimated is the reluctance of national Treasuries and Finance departments to make budget and expenditure procedures more flexible – often for the good reason that the current (inflexible) system achieves effective control of expenditure, a macroeconomic goal that they may (again, quite reasonably) not wish to jeopardise.

Structural Change and "Big Push" Efficiency Improvements

Eliminating very pronounced inefficiencies may well require concerted, deliberately planned structural change. The clearest single example of such a "big push" approach lies in the scope for radical restructuring of hospital provision which is so clearly visible in much of the developing country data. It is one thing to earmark 30-40% of a country's beds and hospitals for closure or restructuring, but quite another to actually achieve the object. Substantial analytical and planning effort will be required, while significant additional funding will be required for implementation. Key areas requiring funding include redundancy payments for retrenched staff; capital costs of site closure and disposal (which can be significant); increased expenditure on professional management; improvement works to upgrade facilities which are remaining open; and, quite possibly, the odd new hospital to sugar the pill for all concerned. While substantial bed closures could be achieved in most developing countries with little problem due to low occupancy rates, in more efficient systems an increasing requirement for the closure of acute beds to be directly offset by additional capacity in nursing home or step-down facilities should also be expected. For a discussion of the one-for-one substitution of UK NHS beds by nursing home beds, see Hensher, Fulop *et al* (1999). None of this detracts from the ultimate realisation of efficiency savings – but it highlights that plans for grand structural change which do not receive adequate funding will most likely go badly awry. The provision of such capital transformation funding would seem to be an ideal use of donor funding – a discrete, non-recurrent programme whose explicit aim is to unlock efficiency savings.

Shifting to Continuous Productivity Improvement

Important as the gains from such putative structural transformation may be, there is a strong case to say that achieving a more modest shift to an organisational culture which expects and delivers year on year productivity improvements is probably more

important still. Certainly, a country which invested heavily in structural change and then did not ensure that efficiency improvement became institutionalised would probably have wasted its money.

In general terms, developed countries have consistently improved productivity in health care over a long time span (Hensher, Edwards and Stokes, 1999). Yet on the basis of impression and fragmentary data (e.g. Zere, unpublished), it seems likely that many developing countries have faced either static or negative productivity and efficiency change over recent years. A number of factors have probably contributed to this lack of demonstrated efficiency gain. Foremost has been a general insufficiency of funds, leading to the bottlenecks and inefficient input mixes described earlier. These are routinely exacerbated by indiscriminate cost-cutting exercises. But I would argue that another key contributor has been the continuing failure to develop a strong cadre of non-medical, professional health service managers in most developing countries. The continued dominance of medically qualified administrators, often with very little or no management training, loath to take on their colleagues, and often still practicing clinical medicine for much of their working day, represents a lost opportunity to spark (or, if necessary, to bludgeon) change.

> "Basically, productivity improvements and cost reduction are normal functions of management. It is axiomatic that the unique function of management is innovation: doing things better; trying to figure out better ways to produce better products at a lower cost, and also to manage an enterprise or an organisation efficiently within a given technology." (Kendrick *et al*, 1981)

Professional managers, armed with simple data with which to benchmark and compare performance, given basic authority to adjust resource use and production processes, themselves judged significantly upon their ability to improve efficiency, supported by what Herrick and Kindleberger (1983) call "…institutions dedicated to the process of change" – this would represent a truly fundamental change in the commitment of health systems in developing countries to improving efficiency.

References

Aletras VH. A comparison of hospital scale effects in short-run and long-run cost functions. *Health Economics* 1999; 8:521-530.

Appleby J. Promoting efficiency in the NHS: problems with the labour productivity index. *British Medical Journal* 1996; 313:1319-1321.

Ashby J, Guterman S, Greene T. An analysis of hospital productivity and product change. *Health Affairs* 2000; 19(5):197-205.

Assembly of Behavioural and Social Sciences Panel to Review Productivity Statistics. Measurement and interpretation of productivity. Washington, National Academy of Sciences 1979.

Barnum H, Kutzin J. Public hospitals in developing countries: resource use, cost, financing. Washington, The World Bank 1993.

Baumol WJ. Economic theory and operations analysis (Fourth Edition). New Jersey, Prentice-Hall 1977.

Berman P. Selective primary health care: is efficient sufficient? *Social Science and Medicine* 1982; 16(10):1054-1059.

Bitran R. Efficiency and quality in the public and private sectors in Senegal. *Health Policy and Planning* 1995; 10(3):271-283.

Breyer F. The specification of a hospital cost function: a comment on the recent literature. *Journal of Health Economics* 1987; 6:147-157.

Briggs AH, O'Brien B. The death of cost-minimisation analysis? *Health Economics Letters* 2000; 4(4):3-10.

Centre for Reviews and Dissemination. Effective Health Care Bulletin: Hospital volume and health care outcomes, costs and patient access. Nuffield Institute for Health, University of Leeds, and NHS Centre for Reviews and Dissemination, University of York 1996.

Chernichovsky D. Allocation of manpower and economic efficiency in medical care. *In* Griffiths A, Bankowski Z. (Eds) Economics and health policy: proceedings of the XIIIth Round Table Conference. Geneva, Council for International Organizations of Medical Sciences 1980.

Cyert RM, March JG. A behavioural theory of the firm. New Jersey, Prentice-Hall 1963.

Daviaud E. Assessing costs of PHC services in an urban district. Cape Town, University of Cape Town Department of Community Health 1999.

Donaldson C. Commentary: possible road to efficiency in the health service. *British Medical Journal* 1994; 309:784-785.

Drummond MF, O'Brien B, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes (Second Edition). Oxford, Oxford University Press 1997.

Edwards N, Hensher M, Werneke U. Changing hospital systems. *In* Saltman RB, Figueras J, Sakellarides C (Eds). Critical challenges for health care reform in Europe. Buckingham, Open University Press 1998.

Ersoy KE, Kavuncubasi S, Ozcan YA, Harris JM. Technical efficiencies of Turkish hospitals: DEA approach. *Journal of Medical Systems* 1997; 21(2):67-74.

Färe R, Grosskopf S, Lindgren B, Poullier J. Productivity growth in health care delivery. *Medical Care* 1997; 35(4):354-366.

Flessa S. The costs of hospital services: a case study of Evangelical Lutheran Church hospitals in Tanzania. *Health Policy and Planning* 1998; 13(4):397-407.

Folland S, Goodman AC, Stano M. The economics of health and health care. New Jersey, Prentice Hall 1997.

Foster S. Supply and use of essential drugs in Sub-Saharan Africa: some issue and possible solutions. *Social Science and Medicine* 1991; 23(11):1201-1218.

Gerdtham U-G, Löthgren M, Tambour M, Rehnberg C. Internal markets and health care efficiency: a multiple-output stochastic frontier analysis. *Health Economics* 1999; 8:151-164.

Ghana Health Assessment Project Team. A quantitative method of assessing the health impact of different diseases in less developed countries. *International Journal of Epidemiology* 1981; 10(1):73-80.

Gilson L. Management and health care reform in Sub-Saharan Africa. *Social Science and Medicine* 1995; 40(5):695-710.

Gilson L, Mills A. Health sector reforms in Sub-Saharan Africa: lessons of the last ten years. *Health Policy* 1995; 32:215-243.

Giuffrida A, Gravelle H, Sutton M. Efficiency and administrative costs in primary care. *Journal of Health Economics* 2000; 19:983-1006.

Grout PA, Jenkins A, Propper C. Benchmarking and incentives in the NHS. London, Office of Health Economics 2000.

Grytten J, Rongen G. Efficiency in provision of public dental services in Norway. *Community Dentistry and Oral Epidemiology* 2000; 28:170-176.

Hensher M. The rationalization and management of hospitals. *In* Feachem Z, Hensher M, Rose L (Eds). Implementing health sector reform in Central Asia. Washington, The World Bank 1998.

Hensher M, Edwards N. Hospital provision, activity and productivity in England since the 1980s. *British Medical Journal* 1999; 319:911-914.

Hensher M, Edwards N, Stokes R. International trends in the provision and utilisation of hospital care. *British Medical Journal* 1999; 319:845-848.

Hensher M, Fulop N, Coast J, Jefferys E. Better out than in? Alternatives to acute hospital care. *British Medical Journal* 1999; 319:1127-1130.

Herrick B, Kindleberger CP. Economic Development (Fourth Edition). Tokyo, McGraw-Hill 1983.

Heyne P. Micoreconomics. New York, Macmillan 1994.

Holloway J, Lewis J, Mallory G. Performance measurement and evaluation. London, Sage 1995.

Holvad T, Hougaard JL. Measuring technical input efficiency for similar production units: 80 Danish hospitals. Firenze, European University Institute 1993.

Hurley J, Birch S, Eyles J. Geographically decentralized planning and management in health care: some informational issues and their implications for efficiency. *Social Science and Medicine* 1995; 41(1):3-11.

Irvin G. Modern cost-benefit methods. 1978.

Kendrick JW, Buehler VM, Shetty YK. Background and overview of productivity improvement programs. New York, AMACOM 1981.

Kleczowski BM. Technological imperatives and economic efficiency in health care. *In* Griffiths A, Bankowski Z. (Eds) Economics and health policy: proceedings of the XIIIth Round Table Conference. Geneva, Council for International Organizations of Medical Sciences 1980.

Knox Lovell CA, Schmidt P. A comparison of alternative approaches to the measurement of productive efficiency. *In* Dogramaci A, Färe R (Eds). Applications of modern production theory: efficiency and productivity. Boston, Kluwer Academic Publishers 1988.

Lancaster K. Introduction to modern microeconomics. Chicago, Rand McNally International 1969.

Lee K, Mills A. Policy-making and planning in the health sector. London, Croom-Helm 1985.

Liebenstein H. Allocative efficiency vs. 'X-efficiency'. *American Economic Review* 1966; 56:392-415.

Lindsay CM. Applied price theory. New York, Holt-Saunders 1982.

Lipsey RG, Chrystal KA. An introduction to positive economics (Eight Edition). Oxford, Oxford University Press 1995.

Liu X. Improving allocative efficiency: a search for policy tools. Geneva, World Health Organisation, forthcoming.

Mahoney TA. Productivity: problems and prospects. New York, Work in America Institute 1980.

Martineau T, Martinez J. Human resources and health sector reform: priorities for analysis and research. *In* Martinez J, Martineau T (Eds). Workshop on human resources and health sector reforms. Liverpool, Liverpool School of Tropical Medicine International Health Division 1996.

McGuire A. The measurement of hospital efficiency. *Social Science and Medicine* 1987; 24(9):719-724.

Meier GM. Leading issues in economic development (Sixth Edition). New York, Oxford University Press 1995.

Mills A. (1990a) The economics of hospitals in developing countries. Part I: expenditure patterns. *Health Policy and Planning* 1990; 5(2):107-117.

Mills A. (1990b) The economics of hospitals in developing countries. Part II: costs and sources of income. *Health Policy and Planning* 1990; 5(3):203-218.

Nyman JA, Bricker DL, Link D. Technical efficiency in nursing homes. *Medical Care* 1990; 28(6):541-551.

Overdyk FJ, Harvey SC, Fishman RL, Shippey F. Successful strategies for improving operating room efficiency at academic institutions. *Anaesthesia and Analgesia* 1998; 86:896-906.

Ozcan YA. Sensitivity analysis of hospital efficiency under alternative output/input and peer groups: a review. *International Journal of Knowledge Transfer and Utilization* 1992; 5(4):1-29.

Ozcan YA, Cotter JJ. An assessment of efficiency of area agencies on aging in Virginia through data envelopment analysis. *The Gerontologist* 1994; 34(3):363-370.

Ozcan YA. Physician benchmarking: measuring variation in practice behaviour in treatment of otitis media. *Health Care Management Science* 1998; 1:5-17.

Palmer S, Torgerson DJ. Definitions of efficiency. *British Medical Journal* 1999; 318:1136.

Parker D, Newbrander W. Tackling wastage and inefficiency in the health sector. *World Health Forum* 1994; 15:107-113.

Peters DH, Elmendorf AE, Kandola K, Chelleras G. Benchmarks for health expenditures, services and outcomes in Africa during the 1990s. *Bulletin of the World Health Organisation* 2000; 78(6):761-769.

Posnett J. Is bigger better? Concentration in the provision of secondary care. *British Medical Journal* 1999; 319:1063-1065.

Posnett J. Volume, outcome, efficiency and access. *In* McKee M, Healy J (Eds). The role of the hospital in a changing environment. World Health Organisation, In Press.

Rice T. The economics of health reconsidered. Chicago, Health Administration Press, 1998.

Rosenman R, Siddarthan K, Ahern M. Output efficiency of health maintenance organizations in Florida. Unpublished.

Rosko MD, Chilingerian JA. Estimating hospital inefficiency: does case mix matter? *Journal of Medical Systems* 1999; 24(2):57-71.

Samuelson PA, Nordhaus WD. Economics (Fifteenth Edition). New York, McGraw Hill 1995.

Simon HA. Theories of decision-making in economics. *American Economic Review* 1959; 49.

Somanathan A, Hanson K, Dorabawila T, Perera B. Operating efficiency in public health sector facilities in Sri Lanka: measurement and institutional determinants of performance. Small Applied Research Paper No. 12. Bethesda, MD, Partnerships for Health Reform Project 2000.

Squire L, van der Tak H. Economic analysis of projects. 1975.

Stigler GJ. The Xistence of X-efficiency. *American Economic Review* 1976; 66:213-216.

Stiglitz JE. The causes and consequences of the dependence of quality on price. *Journal of Economic Literature* 1987; 25:1-48.

Street A, Haycock J. The economic consequences of reorganizing hospital services in Bishkek, Kyrgyzstan. *Health Economics* 1999; 8:53-64.

Todaro MP. Economics for a developing world: an introduction to principles, problems and policies for development. London, Longman 1977.

Tyler LH, Ozcan YA, Wogen SE. Mental health case management and technical efficiency. *Journal of Medical Systems* 1995; 19(5):413-423.

Valdmanis VG. Ownership and technical efficiency of hospitals. *Medical Care* 1990; 28(6):552-561.

White KR, Ozcan YA. Church ownership and hospital efficiency. *Hospital and Health Services Administration* 1996; 41(3):297-310.

Williams A. Measuring the effectiveness of health care systems *in* Perlman M (Ed), The economics of health and medical care. Basingstoke, Macmillian 1986.

Williamson OE. The economics of discretionary behaviour. New Jersey, Prentice-Hall 1963.

World Bank. World Development Report 1993: Investing in Health. The World Bank, Oxford 1993.

World Health Organisation. World Health Report 2000. Geneva, WHO 2000.

Yellen JL. Efficiency wage models of unemployment. *American Economic Review* 1984; 74:200-205.

Zere E, McIntyre D, Addison T. Technical efficiency and productivity of public sector hospitals in three South African provinces. Paper submitted to *South African Journal of Economics*.