Hospital Performance Including Quality: Creating Economic Incentives Consistent with Evidence-Based Medicine

By

Simon Eckermann

A thesis submitted in fulfilment of the requirements of the Degree of Doctor of Philosophy

In

The School of Economics

Faculty of Commerce and Economics

The University of New South Wales

2004

Acknowledgements

I would like to thank Dr. Kevin Fox for his supervision, support, feedback and advice and Knox Lovell for his initial supervision while visiting the UNSW. I would also like to thank colleagues for helpful feedback provided in response to papers presented at: The 2001 and 2002 Australian Health Economics Society Conferences; The economic measurement group workshop '02 at UNSW; The 2002 Australian PHD conference in Economics and Business in Canberra; The 2003 International Health Economics Association Conference in San Francisco; The 2003 International Society for Health Technology Assessment Conference in Canmore, Canada; McMaster University Grand Rounds 27th June 2003, Hamilton; Invited Seminar at the Centre For Health Economics, York University, 17th July 2003, and; the 2003 Australia-New Zealand Health Services Research Conference. Of these colleagues I particularly thank Brita Pekarsky, Karl Claxton, Bernie O'Brien and Tim Coelli for helpful comments. I would like to acknowledge and thank the co-operation of NSW Health in providing and approving the use of 1998-99 DRG level data on hospital costs, admissions and mortality data in illustrating application of the correspondence theorem to relative performance measurement for hospitals. I acknowledge the assistance of and thank Adrienne Kirby for writing SAS code for bootstrapping undertaken in appendix 8.1.

I thank my parents Liz and John for their support and encouragement and my partner Nicola and son Pascal for their patience and understanding in the process of undertaking a doctoral thesis. I dedicate this thesis to the memory of Bernie O'Brien, a friend to whom I owe much for his wisdom and inspiration.

Abstract

This thesis addresses questions of how to incorporate quality of care, represented by disutility-bearing effects such as mortality, morbidity and re-admission, in measuring relative performance of public hospitals. Currently, case-mix funding and performance, measured with costs per case-mix adjusted separation, hold hospitals accountable for costs, but not effects, of care, creating economic incentives for quality of care minimising cost per admission.

To allow an appropriate trade-off between the value and cost of quality of care a correspondence is demonstrated between maximising net benefit and minimising costs plus decision makers' value of disutility events, where effects of care can be represented by disutility events and hospitals face a common comparator. Applying this correspondence to performance measurement, frontier methods specifying disutility events as inputs are illustrated to have distinct advantages over output specifications, allowing estimation of:

1. economic efficiency conditional on the value of avoiding disutility events.
2. technical, scale and congestion sources of net benefit efficiency;
3. best practice peers over potential decision makers' value of quality; and
4. industry shadow price of avoiding disutility events.

The accountability this performance measurement framework provides for effects and cost of quality of care are also illustrated as the basis for moving from case-mix funding towards a funding mechanism based on maximising net benefit. Links to evidence-based

medicine in health technology assessment are emphasised in illustrating application of the correspondence to comparison of multiple strategies in the cost-disutility plane, where radial properties as shown to provide distinct advantages over comparison in the cost-effectiveness plane.

The identified performance measurement and funding framework allows policy makers to create economic incentives consistent with evidence-based medicine in practice, while avoiding incentives for cream-skimming and cost-shifting. The linear nature of the net benefit correspondence theorem allows simple inclusion of multiple effects of quality, whether expressed as not meeting a standard, functional limitation or disutility directly. In applying the net benefit correspondence theorem to hospitals a clinical activity level is suggested, to allow correspondence conditions to be robustly satisfied in identification of effects with decision analytic methods, adjustment for within DRG risk factors and data linkage to effects beyond separation.

Table of Contents

Chapter 3: Level of aggregation for analysis

Chapter 4: Performance measurement specifying quality with disutility events as an output

Chapter 5: The correspondence theorem: specifying quality with disutility events as inputs

Chapter 6: Including quality of care in funding - creating incentives for net benefit maximisation with a budget

Chapter 7: Policy Implications

Chapter 8: The correspondence theoremin health technology assessment

Chapter 9: Future Research

List of Figures

List of Tables

Appendices

Appendix figures

Appendix tables

References

List of Acronyms and Abbreviations

ABS	Australian Bureau of Statistics
ACDHAC	Australian Commonwealth Department of Health and Aged Care
ACHS	Australian Council of Healthc are Standards
AIHW	Australian Institute of Health and Welfare
CDHAC	Commonwealth Department of Health and Aged Care (Australia)
CEA	cost effectiveness acceptance (curve)
CRS	constant returns to scale
DALY	disability adjusted life year
DEA	data envelopment analysis
DMU	decision making unit
DRG	diagnostic related group
HIC	Health Insurance Commission (Australia)
HTA	health technology assessment
ICER	incremental cost effectiveness ratio
IES	inlier equivalent separation
LIPID	Long-Term Intervention with Pravastatin in Ischaemic Disease (clinical trial)
MDC	major disease classification
NEB	net effect benefit
NHCDC	National Hospital Cost Data Committee (Australia)
NHS	National Health Service (UK)
NHCPI	National Health Care Purchasing Institute (USA)
NHPC	National Health Performance Committee (Australia)
NICE	National Institute for Clinical Excellence (UK)
NMB	net monetary benefit
NOAA	National Oceanic and Atmospheric Administration (USA)
PBAC	Pharmaceutical Benefits Advisory Committee (Australia)
PPF	production possibility frontier
QALY	quality adjusted life year
SFA	stochastic frontier analysis
TFP	total factor productivity
VRS	variable returns to scale
WHO	World Health Organisation
WIES	weighted inlier equivalent separation
WTA	willingness to accept
WTP	willingness to pay

Chapter 1: Introduction

1.1 Objectives

This thesis focuses on allowing for quality of care in measuring relative performance and funding of public hospitals inpatient care. The policy objective is to allow incentives for appropriate quality of care in identifying peers, benchmarking and funding. The central economic question addressed is how to incorporate disutility events as health related quality of care indicators into performance measurement and funding mechanisms, to provide appropriate incentives for quality of care. In identifying robust performance measurement and funding mechanisms, questions are also addressed in relation to: the appropriate level of aggregation for analysis; allowing for patient risk factors in production of health care to avoid incentives for cream-skimming and; allowing for the effects of hospital care across the health system to avert incentives for cost-shifting.

The problems and research agenda which motivate this thesis are not new. In Eckermann (1994), problems of measuring and funding based on cost per case-mix adjusted separation, including perverse incentives for lowering quality of care and cost-shifting, were outlined and a research agenda to address these problems proposed. The research agenda suggested the need to:

(1) appropriately reflect the twin objectives for hospitals of minimising costs and maximising health gain, following Harris (1977) and;

(2) allow for the effects of hospitals within a health care system, following Evans's (1981) notion of incomplete vertical integration.

1.2 Overview

The research in this thesis identifies, and illustrates, a framework for performance measurement and funding of inpatient care in public hospitals to create incentives for health related quality of care consistent with evidence-based medicine. In including health related quality of care in performance measurement, an input specification of disutility events valued at decision makers' threshold of willingness to pay is illustrated to have a correspondence with maximising net benefit. An underlying objective of net

benefit maximisation allows the non-tradable and incremental nature of health effects of care to be explicitly addressed. In comparison, performance measured with output specifications of effects of care, at best, correspond to an objective of minimising average cost effectiveness, which fails to account for the incremental or non-traded nature of effects of care.

In chapter 1 the objectives and characteristics of public hospital inpatient care, and transaction conditions under which care takes place, are outlined as a background to the role of performance measurement and funding mechanisms. Current funding mechanisms for inpatient care of case-mix payments and performance measurement, based on cost per case-mix adjusted separation for hospital inpatient care, are introduced. Questions are raised in relation to:

(1) the appropriateness of the implicit objective function underlying these mechanisms and measures;
(2) the incentives created for quality of care;
(3) the impact incentives for quality of care have on other sectors and;
(4) their appropriateness and usefulness as a policy tool at a hospital level of aggregation.

In chapter 2, failures of current measures to reflect the objective of improving health or place hospitals within the health care system are demonstrated to result in incentives for cost minimising quality of care and cost-shifting to care beyond-separation. Incentives for appropriate quality of care are suggested to require the objective function to reflect an appropriate trade-off between health effects and cost of quality of care. Lack of accountability for quality of care creates a performance-efficiency paradox. Hospitals measured as efficient in minimising cost per admission, if due to lower quality of care, can have worse health outcomes and greater costs in treating a patient population across a health care system over time. Technical efficiency measured as minimising inputs per admission or economic efficiency measured as minimising costs per admission are both neither necessary nor sufficient for net benefit maximising care.

The comparison of relative performance, identification of peers and benchmarking with measures such as cost per case-mix adjusted separation originating at an aggregate level, is also demonstrated to face a problematic aggregation paradox. A hospital with lower cost per admission in each individual activity can be measured as less efficient at an aggregate level with cost per case-mix adjusted separation, due to exogenously determined cost shares between activities, which case-mix adjustment does not control for.

In chapter 3, measuring performance at a clinical activity (diagnostic related group (DRG)) level is identified as allowing both performance-efficiency and aggregation paradoxes to be overcome, with the ability to:

1. flexibly identify effects of care by activity in reflecting an objective of health maximisation;
2. adjust for differences in patients prognostic factors by DRG across hospitals to avoid cream-skimming incentives;
3. use decision analytic methods to comprehensively identify within- and beyond-separation effects of care, and avoid incentives for cost-shifting;
4. reveal inefficiency at a clinical activity level hidden by aggregation, and;
5. avoid the Fox (1999) aggregation paradox, with non-paradoxical aggregation from weighting clinical activity (DRG) level economic efficiency with standardised industry cost shares.

The question of how to include disutility events in performance measurement to reflect an appropriate objective function and create appropriate incentives for quality remains.

Chapter 4 illustrates problems with attempting to measure hospital performance under a quality-quantity trade-off where effects of care are specified as outputs.

Output specification of admissions (quantity) and effects (quality) as inverted disutility event are illustrated as able to avoid size biasing if inverted rates are normalised relative

to admissions. However non linearity in valuing disutility events creates dichotomous incentives for low cost, low quality and or high cost, high quality care.

An alternate specification of admissions without disutility events recognises quality as interacting with activity and provides disincentives and incentives for quality of care. However, this specification constrains quality to having an extremely restrictive functional relationship with activity. Admissions with disutility events are valued at 0 regardless of the seriousness or the triviality of disutility events. While this specification is potentially appropriate for disutility events of readmission to the same DRG, a general inability to flexibly trade-off the cost and value of quality remains

The hyperbolic method of Färe, Grosskopf, Lovell and Parsuka (1989), equi-proportionally decreasing weakly disposable negative outputs (such as pollution) and increasing marketed desirable outputs (such as electricity) is shown to be problematic in translating to a setting where disutility event rates represent service quality Regions of the hyperbolic frontier where undesirable outputs increase and desirable outputs decrease (whether with weak or strong disposability of undesirable outputs), are demonstrated as problematic unless undesirable outputs are assumed exogenously determined. Performance measured with technical efficiency relative to this frontier reflects perverse choice of quality of care below a cost minimising level (disutility event rate above a cost minimising level). The inability to represent an output orientated economic efficiency measure under the hyperbolic approach implies these perverse choices cannot be subsumed into allocative inefficiency.

Congestion efficiency as the residual of technical efficiency under strong (costless), and weak disposability of undesirable outputs, is also shown to lack a meaningful interpretation, where disutility event rates represent health effects of quality of services. Where outputs are separable, such congestion inefficiency can be interpreted as potential additional production, of a desirable output, possible in the absence of regulation or

restriction on the undesirable output (such as pollution)[1]. However, in hospitals, congestion inefficiency as a measure of increased admissions, fails to recognise the derived nature of demand for admissions in ignoring effects of care. Finally, in the absence of market prices for a separable 'desirable' output, a monetary shadow price of avoiding disutility events is also not estimable with an output specification of disutility events using the method of Färe, Grosskopf, Lovell and Yaisawarang (1993).

In general output specifications of disutility events are found unable to account for the incremental and non-tradeable nature of health effects, or an appropriate trade-off between incremental value and cost of quality. Hence output specifications of effects, whether framed as utility or disutility bearing rates do not provide appropriate economic incentives for quality of care.

In Chapter 5 a specification of effects within a ratio measure of performance is developed consistent with an underlying objective of net benefit maximisation. Net benefit maximisation trades off the monetary value of incremental health effects less incremental cost of technologies or strategies relative to an appropriate comparator (next best practice).

Maximising net benefit directly does not allow ratio measurement of efficiency as inputs and outputs are not necessarily non-negative. However, a linear transformation is shown to allow a ratio measure corresponding with maximising net benefit per admission where effects of care are framed as disutility event rates and specified as inputs in performance measurement, under correspondence conditions of coverage and comparability.

In any bilateral comparisons between hospitals, maximising net benefit per admission has a one-to-one correspondence with minimising cost per admission plus the decision makers' monetary value of avoiding disutility per admission where:

[1] Although it is also arguable whether this interpretation is appropriate given the implicit 0 value ascribed to undesirable outputs (e.g. pollution) and as economic efficiency is not calculable under the hyperbolic approach.

(1) hospitals can be compared as if they face the same (unknown) comparator (differences in expected costs and disutility event rates are adjusted for) and;

(2) relative effects of care are covered by disutility event rates.

This correspondence provides a general method and framework (in satisfying correspondence assumptions) for relative performance measurement, consistent with an underlying objective of net benefit maximisation, and hence evidence-based medicine. Application of this correspondence to measuring relative hospital performance at a clinical activity level is illustrated using the frontier method of data envelopment analysis, including effects framed as disutility events as inputs, and with admissions as a sole output. Economic efficiency measured under this specification with disutility events priced at decision maker's value (willingness to pay threshold), corresponds with net benefit maximisation, under correspondence conditions of comparability and coverage. Technical efficiency relative to a frontier minimising cost (ideally physical inputs) and disutility events per admission, appropriately reflects dual hospital objectives of health maximisation and cost minimisation (Harris, 1977). Unlike cost efficiency in production of admissions *per se*, ignoring effects of quality of care, this technical efficiency measure is shown to be necessary for maximising net benefit per admission.

Where decision maker's threshold value of avoiding disutility events is unknown, regions over which potential peers are best practice are illustrated to be simply identified by back-solving between adjacent technically efficient hospitals on the frontier. The shadow price of quality implicit in industry behaviour is shown as easily estimated and interpreted as the value of avoiding disutility events where industry (the cost-share weighted sum of each hospital's) economic efficiency, or allocative efficiency, are maximised.

Satisfying correspondence conditions of a common comparator and coverage of effects of care, while allowing net benefit maximising quality of care, are also demonstrated to provide a framework for avoiding incentives of cost-shifting and cream-skimming.

Satisfying coverage of effects by disutility events is facilitated at a clinical activity level with the ability to use decision analytic methods in identifying effects of care. Perceived utility bearing effects of quality of care can also be reframed as disutility event rates where they can be expressed as either:

(a) health related standards of physical, mental or social functioning;

(b) cardinal measures of functional ability that are measurable at point of separation, or;

(c) health related utility.

The linear nature of the correspondence allows easy inclusion of patient populations with combinations of multiple disutility events, as well as effects beyond-separation. The ability to reframe utility bearing aspects of quality as disutility events, and include multiple disutility events in this linear framework suggests there is no technical barrier to coverage of effects of care in avoiding incentives for cost-shifting or quality-skimping. At a clinical activity (DRG) level decision analytic methods can be employed to be comprehensive in coverage of effects and in identification of risk factors.

Methods are identified to adjust for differences in patient population risk factors within-DRG in satisfying the common comparator assumption in practice to avoid incentives for cream-skimming.

Chapter 6 shows how the performance measurement framework identified in chapter 5 can be extended to a funding mechanism for inpatient care in public hospitals at a clinical activity (DRG) level. Where correspondence conditions are satisfied conditioning payments on the value of measured effects of quality of care creates hospital accountability for their effects relative to cost of quality of care. A sequential two-stage funding is illustrated that allows a planned and managed process in systematically moving economic incentives for quality of care from those of minimising cost per admission under current case-mix funding towards those of net benefit maximisation, while remaining within case-mix funding.

Funding hospital inpatient activities using prospective case-mix payments based on average expected cost for each clinical activity (DRG) alone are shown to implicitly value effects of care at 0 creating incentives for quality minimising cost per admission. In practice, by failing to hold hospitals accountable for the value of quality of care the funding of each inpatient activity at the average cost of providers with variable quality of care, also allows technical inefficiency to be hidden behind lower quality of care. Hospitals providing quality of care at a cost minimising level are not required to have cost minimising cost of care, where case-mix funding is based on average industry expected costs.

To allow a managed transition in payments reflecting net benefit maximising value for quality of care, a sequential two-part funding mechanism is developed and illustrated. The first-part payment conditions relative funding between hospitals on quality of care relative to expected cost, at best practice, for a scheduled value of quality. The second-part fixed payment per admission acts as a buffer and stopping mechanism for value change, providing payments to equalise a fixed (for example existing case-mix) budget per admission. The proposed mechanism sequentially shifts value of quality from current industry shadow price, as identified in chapter 5, towards the decision maker's value. This allows policy makers to remain within existing case-mix funding per admission while incrementally shifting incentives, under correspondence conditions, towards economic incentives for quality of care maximising net benefit per admission. Economic incentives are created for technical and allocative inefficiency to be converted into improved quality, as hospitals become accountable for the value as well as the costs of quality of care.

Chapter 7 outlines policy implications from proposed performance measurement and funding mechanisms of hospitals.

Problems of current hospital performance measures and funding mechanisms failing to reflect objectives of hospitals, and consequently providing incentives to reduce quality of care, have been recognised by policy makers. Applying the correspondence theorem to

performance measurement and funding, allows an objective function consistent with maximising net benefit and consequently an appropriate trade-off between cost and value of quality of care. Satisfying correspondence conditions of a common comparator and coverage of effects of care also provides a framework for avoiding cream-skimming and cost-shifting in incorporating quality of care consistent with net benefit maximisation. For correspondence theorem assumptions to be satisfied and appropriate incentives to be created, existing policy agendas remain in:

1. adjusting for prognostic factors within DRG to prevent cream-skimming and satisfy the common comparator assumption, and;
2. use of either data linkage to effects attributable to care informed by decision analytic methods, or measurement of average health status at point of separation, to prevent incentives for cost and event shifting and satisfy the coverage of effects assumption.

At a clinical activity (DRG) level, decision analytic methods allow patient risk factors and appropriate effects of care to be identified for incorporation in the performance measurement and funding framework provided by the correspondence theorem. Clinical audit and peer review processes can be utilised in monitoring and verifying reported effects of care within hospital at a clinical level.

Where correspondence theorem requirements are satisfied, policy implications of creating economic incentives for hospitals, consistent with net benefit maximisation and avoiding cream skimming and cost-shifting, are significant. Hospitals with these incentives become economic agents for evidence-based medicine in their choice and use of technology in practice. With appropriate data linkage, informed by decision analytic methods, this can be extended to referral practices. Hospitals are economically accountable for the effects and costs of quality of care consistent with net benefit maximisation and evidence-based medicine in practice. In considering the internal relationship between clinicians and administrators within hospitals (following Harris, 1977), the proposed performance measurement and funding framework shifts the implicit objective function of the administrator from cost minimisation to net benefit

maximisation. The mechanism values, but also holds clinicians accountable for effects of quality of care, while hospital administrators can no longer act as accountants minimising costs per admission but are required to trade off the value against cost of care. Under correspondence conditions, this allows administrators and clinicians objectives, as well as the funding body as principal, and hospital as agent, to be more closely aligned. This should reduce tension between clinicians and administrators, facilitating a negotiation process within hospitals akin to program budgeting and marginal analysis (PBMA). Explicit and more appropriate incentives for quality of care, corresponding with health care objectives by activity, are created. This reduces reliance on partial performance measures, localised incentives from clinical standards or regulation, and local negotiation conditions for cost minimisation versus health maximisation, within-hospital and by clinical activity. Unlike current reliance on local conditions and regulation, economic incentives would be uniform and continuous.

Chapter 8 illustrates application of the correspondence theorem to relative comparison of strategies in the cost disutility plane for health technology assessment. In health technology assessment, where multiple strategies are compared, frontiers in cost-disutility space are illustrated to have advantages over frontiers in the incremental cost-effectiveness plane (which do not allow radial contraction). The correspondence condition of a common comparator is shown to be naturally satisfied by randomised control trial evidence. Disutility event rates are easily constructed to allow coverage of effects. Rates of survival or morbidity avoided, as an incremental effect, are naturally reframed as mortality and morbidity rates. Life years or QALYs saved can be translated to life years or QALYs lost relative to the strategy with greatest average effectiveness.

In the cost disutility plane the frontier of best practice strategies is easily and intuitively identified and characterised, as are dominance and extended dominance, differences in net benefit and regions of WTP over which strategies are preferred. Distinct advantages over comparison in the incremental cost- effectiveness plane include the ability to estimate degree of dominance (as technical inefficiency) and its degree of uncertainty, and represent differences in net monetary or effectiveness benefit as distances between

isocost curves on the cost and disutility axis. In allowing for uncertainty, , a method for improving precision in bootstrapping incremental cost effectiveness ratios, and probability of dominance from randomised control trial patient level data, is identified and illustrated.

Chapter 9 discusses the robustness of the correspondence theorem framework and considers future research in allowing for uncertainty and applications of the correspondence theorem to performance measurement allowing for quality of care, in other settings.

In allowing for the effect of sampling variation on the position of the best practice frontier, future research on use of stochastic frontier analysis and Bayesian shrinkage estimation methods are suggested. To the extent that such research allows underlying frontiers to be estimated removing effects of sampling uncertainty, questions of allocative efficiency between hospital activities could then be more robustly addressed. In funding at an industry level, while relative payments are suggested to be based on observed disutility event rates, a minimum second stage buffer payment in the two stage funding mechanism, illustrated in chapter six, could be based on an estimated underlying, rather than observed, frontier. In funding individual hospitals, while aggregation across individual activities (DRGs) and time reduces the uncertainty of effects, risk sharing between smaller hospitals could be considered. However, in general, if correspondence conditions have been satisfied with adjustment for risk factors and beyond-separation effects, only hospitals that have systematically hidden technical efficiency behind low quality of care need fear accountability for quality of care.

Application of the general correspondence theorem method and framework to relative performance measurement and funding is suggested, where net benefit maximisation is an appropriate objective, and quality of service can be represented by reduction in rates of disutility events. These industries include other health care and service industry settings such as employment placement and corrective services. In these industries effects from quality of services are incremental and non- tradable (and hence average

cost-effectiveness is not an appropriate objective) and adjustment for risk factors and effects beyond-service are important. Application of the correspondence theorem is also suggested as potentially valuable in private sector service industries, where prices do not reflect quality due to transaction conditions, and quality can be represented by disutility events.

Chapter 10 concludes. Questions of specifying effects of service quality with an appropriate underlying objective function, the level of analysis and placement of hospitals within the health care system, in performance measurement and funding, are addressed.

1.3 Background: hospital characteristics and objectives

In a perfectly competitive market, firms have an incentive and imperative for efficient production in order to be competitive with other providers of a homogenous good or service, given consumers with perfect information on prices of competitors. In the health care sector, care is quality differentiated and patients frequently have little (often misleading) information or experience of type or quality of health care provided. This is particularly the case for patients in hospitals, where increased complexity in decision making for prognosis, diagnosis and treatment all but ensure bounded rationality (Simon, 1957) in distinguishing between providers' quality of care. Because each patient is potentially different, complexity in decision making can also be compounded by 'small numbers problems' (Arrow, 1969), requiring customisation of care to patient characteristics (risk factors).

While patients value quality of care to improve their health outcomes, they face transaction conditions and information constraints such that they are unable to distinguish between provider quality of care. This inability of patients to judge quality of care is likely to exist, both ex-ante and ex-post (McGuire, Henderson and Mooney, 1988:43-44), given patients have difficulty in specifying 'counter-factual' outcomes in the absence of alternative care provided (Weisbrod, 1978:52). As McGuire (1987:170) argues, in outlining the requirement for the agency role of providers in hospitals, consumers rely on

information given by the provider on the relationship between health care, health status and expected outcome.

The gravity of making potentially wrong decisions also leads to this requirement for agency. As Weisbrod (1991) noted, hospital services differ from other commodities because technical complexity of the commodity is a base for uncertainty and informational asymmetry, but also because services can affect preservation or quality of life. There are, therefore, high costs associated with both decision making and wrong decisions.

In the terms of the transaction theory framework of Williamson (1975), patients face bounded rationality (Simon (1957)), information impactedness, uncertainty and complexity in decision making. Under these transaction conditions, providers of hospital services lack natural economic imperatives, or incentives, to provide health related quality of care efficiently.

In undertaking relative economic efficiency assessment, integration of the value of a homogenous good or service is normally assumed to be represented by the market price. However, in hospitals, transaction conditions are not present for prices of admission to reflect quality of care, even if they were provided in a market. The role of performance measurement and funding for public hospitals can, therefore, be seen as attempting to create appropriate economic incentives for health care provision, providing what Donaldson and Gerard (1993) term the 'visible hand'. For this visible hand to provide appropriate incentives for efficient provision of hospital care, consideration of the quality differentiated nature of care and its effects on outcomes of care, is required.

The problem, then, is how to fund and measure the economic performance of hospitals within the health sector, to create appropriate incentives for efficient provision of health care. In creating appropriate incentives, Smith (1995) and Goddard, Mannion and Smith (2000) suggest that the measurement of performance, attribution of performance and reward for performance are all important aspects. In a principal-agent framework the

principal's (funder's) problem is characterised as designing a reward system based on measured outcomes to maximise utility from health outcome relative to reward paid to agents (hospitals as providers) in treating a patient population.

The value of effects from quality of care, as well as implicitly the cost or resource use of quality of care, needs to be included in performance measurement to reflect an appropriate economic objective, and create appropriate incentives. As Lovell (1993) noted in identifying economic efficiency for appropriate performance measurement:

> "It is also possible to define the optimum in terms of the behavioural goal of the production unit. In this event, efficiency is economic and is measured by comparing observed optimum cost, revenue, profit or whatever the production unit is assumed to pursue, subject of course to appropriate constraints on quantities and prices." Lovell (1993:4).

Economic efficiency measures which do not represent an appropriate objective function, or only reflect a partial objective function or selected variables, can be problematic. Partial performance indicators create incentives to comply with measured indicators at the expense of other objectives (Smith, 1995). As Stigler (1976:213-14) observed, measured inefficiency may be a reflection of a failure to measure the right variables, constraints or economic objective.

In hospitals, quality of care affects both health outcomes and resource use, or costs of care. To create appropriate incentives for quality of care in efficiency measurement requires inclusion of the value of health effects, as well as the costs of quality, in the implicit objective function. Otherwise, if peer identification or benchmarking ignores the health effects of care, then hospitals are provided with incentives to reduce health related quality of care, while this reduces expected costs.

1.4 Current hospital performance measurement and funding

In Australian hospitals the current measure of output is admissions, weighted by factors reflecting relative expected cost by type of admission. This weighting process is called case-mix adjustment.[2] Relative weights are constructed from expected costs for inpatient admission by diagnosis, referred to as Diagnostic Related Groups (DRGs). In Australia, Australian National Diagnostic Related Group (AN-DRG) weights are estimated each year for more than 660 AN-DRGs by hospital type[3] in the National Hospital Cost Data Collection (CDHA (2000)) sample across public and private hospitals. Economic performance of Australian public hospitals is subsequently measured in comparing inpatient services for purposes of benchmarking and peer identification with cost per case-mix[4] adjusted admission at a hospital level.

Case-mix adjusted admissions are also used as the basis for allocating funding to hospitals. In Australia, in the state of Victoria alone, case-mix funding in 2000/2001 accounted for 70% of a $4 billion health budget. The claims made in support of such case-mix funding are that it:

> "..has enabled hospitals to make more informed decisions on best and most appropriate use of their resources. Case-mix funding encourages more efficient patient treatment and recognises the costs associated with different procedures." (Brook, 2002).

The use of cost per case-mix adjusted separation as a performance measure and funding mechanism, however, raises several questions which need to be addressed in considering such claims:

---

[2] In ABS performance measurement across time, these weights are fixed in a base year.

[3] DRG case-mix weights are separately calculated for public teaching and non-teaching hospitals, major urban and non-major urban hospitals.

[4] Case-mix weights by admission type in Victoria are further adjusted for inpatient stay by an inlier equivalent separation (IES) factor, with a weighting of one for 'normal' length of stay, a higher weighting for longer stay and lower weighting for short stay in adjusting the DRG weight. Long and short stays are defined as over 3 times the average length of stay, maximum 100 days, and less than one third average length of stay, respectively. The IES weights applied are the fraction of one third for lower bound, and one plus days above the upper bound, divided by twice the average length of stay. The IES attempts to capture some of the effects of case-mix within DRG. However, the IES weighting can capture endogenous as well as exogenous effects on length of stay. The product of IES and DRG weight are called weighted inlier equivalent separations (WIES).

1. Is efficiency, measured as the ratio of (cost of) inputs to health service activity as output, an appropriate measure of hospital objectives and performance?
2. Is economic efficiency in production of admissions a necessary condition for efficient production of patient treatment?
3. How appropriate are the incentives created in monitoring performance (identifying peers and benchmarks) or funding on cost per case-mix adjusted separation?
4. How appropriate and useful is efficiency measurement, at an aggregate hospital level, as a policy-making and management tool?

Chapter 2: Problems of performance measurement with cost per case-mix adjusted separation

2.1 Overview

Activity-based and aggregate measures of public hospital performance such as cost per case-mix adjusted separation are demonstrated to face problematic performance-efficiency and aggregation paradoxes.

Measuring public hospital performance with cost per case-mix adjusted separation provides incentives to minimise cost per admission, but creates incentives to reduce quality of care, given that quality of care is not costless. In failing to include value of quality in current performance measurement, a performance-efficiency paradox arises where cost per case-mix adjusted separation can be lowest, but, if due to lower quality of care, this can reflect a high cost per health gain, or low net benefit per admission. As quality of care affects both costs and outcomes of care, cost minimisation conditional on quality of care is a necessary condition for efficient patient treatment, but cost minimisation per admission *per se* is not.

Where multi-product firms have exogenously determined effort across multiple activities recent research has shown that a problematic paradoxical results can occur when economic efficiency scores are derived at an aggregate level. These conditions are shown to be characteristic of public, but not private, hospitals, with current aggregation efficiency measures such as cost per case-mix adjusted separation. This result suggests that comparisons of public hospital economic efficiency should originate at a disaggregated level.

2.2 Case-mix adjustment: increasing homogeneity of what?

Compared to measuring admissions alone, case-mix adjustment of admissions[1] increases the degree of homogeneity in measuring inpatient activity of hospitals, to the extent that case-mix weights represent relative value of outputs across different activities. In estimating average expected costs of treatment, the ability to adjust for case-mix or severity of patient population, with this method, is limited to the extent that

[1] Using relative average costs in a sample of hospitals across diagnostic related groups.

exogenous factors within DRG, such as age and co-morbidities, are not adjusted for. Inpatient activity, adjusted for DRG case-mix weights, clearly goes some way towards adjusting for exogenously determined severity of patient populations, in estimating average expected costs. Therefore, it can mitigate, to some degree, against incentives of hospitals to cream-skim in choice of patients by diagnostic related groups.

However, while case-mix adjustment may move towards the "highest degree of homogeneity" [McGuire 1987:92] in expected cost of inpatient activity, this does not necessarily say anything about relative value of different activities. The assumption that relative average cost between DRGs represents relative value of care, is likely to be seriously flawed, in the absence of a perfectly competitive market to align marginal cost through price to be equivalent to marginal benefit. Arguably, the more important limitation in comparing output of hospitals using cost per case-mix adjusted admission, is that, for each DRG, there is an implicit assumption of homogeneity of quality of care between hospital providers.

Despite a primary objective of health care of improving health in treated patients, and the derived nature of any demand for hospital services, differences in health outcomes from variation in clinical practice or quality of care are ignored with this measure. Therefore, while case-mix adjustment mitigates against cream-skimming incentives to the extent that it increases homogeneity of the expected costs of treated patients, case-mix weights reflect expected costs, not the value of effects from quality of services.

2.3 The performance-efficiency paradox

The primary objective of health care is to improve the health of patients. In providing health care to treat a given patient population, Harris (1977) characterised hospitals as having dual objectives of minimising costs (administrators) and maximising health gain (clinicians). Measuring efficiency as minimising cost per case-mix adjusted separation, while addressing the cost of care in treating a patient population, fails to consider the health effects of care and the value of health related quality of care. Despite this, measurement of production functions and performance have focussed, since Feldstein (1967), on costs and intermediate outputs which have the "highest attainable degree of homogeneity" (McGuire, 1987:92).

Technical or economic efficiency in production of hospital admissions *per se* is, however, neither sufficient, nor necessary, for efficient provision of health care. Technical efficiency in production of admission is only necessary to the extent that the hospital with the lowest cost per health gain (or highest net benefit from value of incremental effectiveness relative to incremental cost of care) has a lower cost per admission, relative to other hospitals with the same quality of care (health gain). However, a low cost per admission *per se*, ignoring quality of care, does not imply a lower cost per health gain, or better performance, than hospitals with higher cost per admission.

Hospitals measured as the most efficient on the basis of resource use or cost per admission, if attributable to lower quality of care than other hospitals, may in fact be the worst performing hospitals. This is particularly the case when considering the effects of hospital activities within their health care systems more broadly. Discretion of providers in relation to point of separation or referral, in combination with incomplete vertical integration between health care sectors (Evans, 1981), implies account needs to be taken of the health and resource effects beyond-hospitalisation. This is particularly the case where there are economic incentives for cost shifting and lower quality, or 'quicker-sicker', care. These effects can include increasing readmission rates to hospitals, treatment in other institutional settings (nursing homes, general practice and specialist services) or in non-institutional settings such as that of informal family care. Health effects, and subsequent treatment and resource follow-on effects, attributable to treatment within hospitals need to be considered in evaluating the performance of hospitals in treating a patient population.

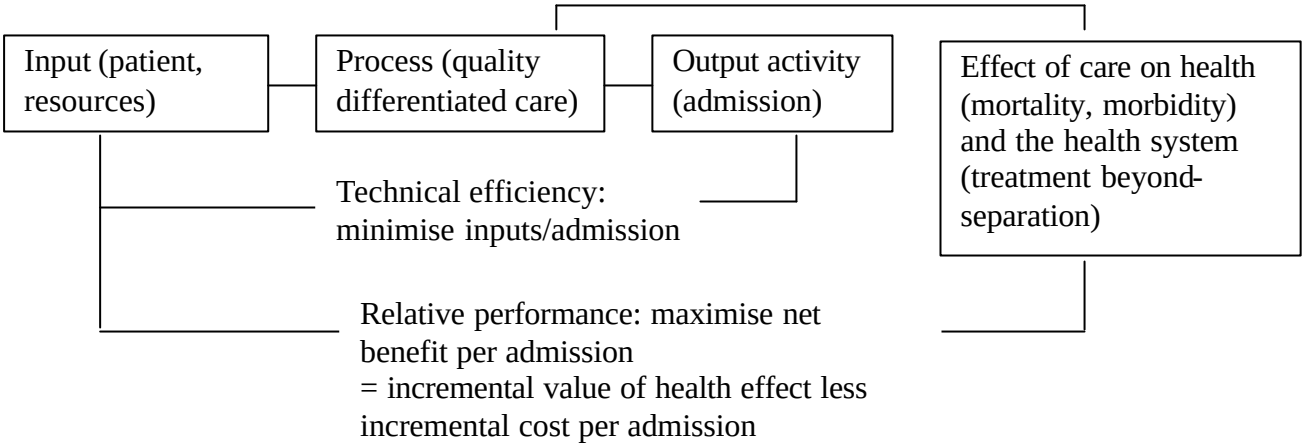
Lower costs attributable to lower quality can even lead to worse health outcomes and higher costs in treating a patient population over time, accounting for effects on the health system beyond-separation. Measuring lowest cost per admission as efficient is clearly problematic where performance can be dominated, on both objectives of health maximisation and cost minimisation, at a health system level. This has lead to calls for including resource use beyond point of discharge in comparing hospital performance (Duckett, 2000). However, including resource effects beyond hospital separation still does not value the health effects of quality of care. For example, a hospital with the highest rate of morbidity and mortality this would still be considered as the most

efficient hospital provided it had the lowest cost per patient including costs beyond admission. The lowest cost of care including costs beyond point of separation can still reflect the lowest quality of care and, potentially, worst performance. More generally, including costs beyond discharge does not prevent incentives for quality skimping where reduced quality of care has non-treatable effects beyond point of discharge.

In measuring relative performance of hospitals, analogous to profit maximisation, maximising net benefit (Stinnett and Mullahy, 1998) per admission, as the value of incremental effects of care less incremental costs (including those beyond-care), provides an appropriate economic objective. Comparing hospital performance with net-benefit per admission, dual objectives of cost minimisation and health maximisation (Harris, 1977) are incorporated under an appropriate trade-off between cost and value of quality. Hospitals are held accountable for both the costs and effects of care, each of which are influenced by quality of care. With an objective of maximising net benefit per admission, even if a hospital has lowest costs within and beyond care, if health related quality of care, and hence expected effects of care, are also lowest, it can still have the lowest net benefit per admission.

The lack of correspondence between performance and economic or technical efficiency, measured as cost or resource use per service (admission), has been described by Pekurinen, Sintonen, Pitaken, Alander and Coyle (1991) as a 'performance-efficiency paradox' in health care generally. As illustrated in figure 2.1, unless health and resource effects beyond-separation of quality of care, as well as implicit within-admission costs of quality of care, are allowed for in efficiency measurement, this 'performance-efficiency paradox' can arise. A hospital with the lowest average cost per case-mix adjusted separation, attributable to lower quality or 'quicker-sicker' care currently measured as most efficient, may represent lowest net benefit or even worse outcomes and potentially highest costs to the health system.

Figure 2.1: Quality of care and the performance-efficiency paradox



The general implication of this performance-efficiency paradox is that whether any strong implications for performance of hospitals can be derived from measuring economic efficiency depends on the ability to incorporate the value of the effects, as well as the costs, of quality of care.

Given a primary objective of improving health, cost efficiency of hospital services, measured in minimising cost per admission (case-mix adjusted or otherwise) is misleading as a measure of hospital performance. As Roos (2002:250) suggests, the use of activity measures such as bed days, treated patients or discharges as proxies or indicators of hospital output is:

> "..very limited because they measure the means to producing real output rather than the output itself".

For cost per case-mix adjusted separation to proxy for performance, assumptions are required that there is a constant quality of care and depth of integration with other health care. These assumptions are required across hospitals in cross-sectional analysis and over time in longitudinal analysis. In ignoring the effects of hospital services, but implicitly including their cost, these assumptions are, however, unlikely to hold, with economic incentives created for hospitals to reduce quality of care and shift events and costs beyond point of separation. An economic incentive is created to reduce quality of

care for any given hospital inpatient activity, while infra-marginal costs of lower quality care across patients outweighs intra-marginal cost from treating additional disutility bearing events within admission.

To avoid these incentives requires that quality of care is incorporated into performance measurement. While technical or economic efficiency do not normally include quality, as Gregan and Bruce (1997:60) state, in using the reciprocal of readmission rates as an output representing quality of care alongside admissions in their study of efficiency of Victorian hospitals:

> "The assumption of the study was that an increase in output using the same quantity of inputs and at least maintaining the same quality standards was a true increase in efficiency, whereas the same increase in output with a fall in quality may not …This is because quality is a defining characteristic ... it is easier to produce lower quality rather than higher quality of output. Therefore, ignoring the quality dimension results in a flawed view of any efficiency results."

A similar view was also expressed by Puig-Junoy (1998:267) in a study of relative performance at a clinical activity level (in peri-natal intensive care):

> "Measurement of efficiency in health services is biased by the way the quality dimension of output is (not) measured."

More generally, as Thanassoulis, Boussofiane and Dyson (1995:590) observed:

> "..where activity levels and quality jointly account for use of resources, both should feature in assessing how efficiently those resources are used."

2.4 The Fox aggregation paradox

In economic efficiency measurement, an aggregation paradox arises at a hospital level of analysis, which is problematic for public hospitals. Public hospitals are multi-product firms (many inpatient and outpatient activities) with largely exogenously determined cost shares by activity, given a requirement to treat patients presenting for care. Under such conditions, a hospital may be more efficient than another in each individual activity, yet be measured with standard economic efficiency measures at an aggregate level as less efficient, due to exogenously determined differences in cost

shares. This result, identified by Fox (1999), is referred to as the 'Fox paradox', following Färe and Grosskopf (2000).

Formally, for a firm with *n* outputs or activities, economic efficiency (*EE*) is conventionally calculated as the ratio of predicted minimum costs to actual costs which can be re-expressed as efficiency in each activity, weighted by cost share of that activity:

$$EE = \hat{TC}/TC = (\sum_{i=1}^{n} \hat{C}_i / \sum_{i=1}^{n} C_i) = s_1 \times E_1 + s_2 \times E_2 + ... + s_n \times E_n \qquad (2.4.1)$$

where:

$\hat{TC}/TC$ is the predicted relative to actual total cost;

$E_i = \hat{C}_i / C_i$ is the economic efficiency in production of activity *i* and;

$s_i = C_i / \sum_{i=1}^{n} C_i$ is the share of total costs of activity *i*.

Hence aggregate economic efficiency measured in a conventional way depends on both economic efficiency in individual activities (varying from 0 to 1) and cost shares by activity. If these cost shares are determined exogenously, then the paradox is problematic as multi-product firms may be rewarded or punished for factors outside of their control. In any bilateral comparison, a hospital can be more efficient than another but be measured as less efficient at an aggregate level, due to exogenously determined cost shares by activity not favouring its comparative advantage.

2.4.1 Illustrating the 'Fox aggregation paradox'

To illustrate the paradox with a simple example, let there be two public hospitals (A and B) from a population of public hospitals producing 2 outputs: inpatient service 1 and inpatient service 2 (table 2.4.1). Let hospital A be more cost efficient than B at producing both inpatient service 1 and inpatient service 2.

Table 2.4.1: The Fox aggregation paradox - a simple hospital example

	Hospital A	*Hospital B*
Economic efficiency (relative to best practice)		
Inpatient service 1	0.9	0.8
Inpatient service 2	0.6	0.5
Cost share		
Inpatient service 1	0.5	0.9
Inpatient service 2	0.5	0.1
Overall economic efficiency	0.75	0.77

In the production of inpatient service 1, let the economic efficiency relative to best practice (however determined) of hospital A (say 90%) be greater than that of hospital B (say 80%). For inpatient service 2, let hospital A again have economic efficiency greater than that of hospital B (say 60% vs. 50%). Now, assume that the exogenously determined relative use of services is such that inpatient service 1 contributes 50% of costs for hospital A and 90% for hospital B (table 2.4.1).

The aggregated economic efficiency measured with the standard definition (cost share weighted sum of economic efficiency in each activity) is 0.75 for hospital A, but 0.77 for hospital B. Hospital A is more efficient in producing each individual activity than B, but is measured as less economically efficient at an aggregate level, due to A's exogenously determined higher relative shares of effort (cost), in activities in which A is relatively less efficient.

The general extension of this simple example, to the aggregate economic efficiency measure of cost per case-mix adjusted separation between hospitals, requires only that relative cost per case-mix adjusted separation for each hospital be viewed as cost efficiency of individual DRG outputs weighted by their cost shares.
That is, (2.4.1) becomes:

$$EE = \hat{TC}/TC = (\sum_{i=1}^{n} \hat{C}^{drg}{}_i / \sum_{i=1}^{n} C^{drg}{}_i) = s^{drg}{}_1 \times E^{drg}{}_1 + \ldots + s^{drg}{}_n \times E^{drg}{}_n \qquad (2.4.2)$$

where for each hospital:

$E^{drg}{}_i = \hat{C}^{drg}{}_i / C^{drg}{}_i$ is cost efficiency in production of diagnostic related group activity i and;

$s^{drg}{}_i = C^{drg}{}_i / \sum_{i=1}^{n} C^{drg}{}_i$ is the share of total costs for diagnostic related group activity i.

Differences in economic efficiency between hospitals cross-sectionally, or across time, can be due to differences of efficiency in individual activities or differences in cost shares between DRGs. Case-mix weights, while standardising on expected costs of admissions by activity, do not standardise on cost shares by activity. As a result, a hospital could be more efficient in providing each inpatient service (DRG) than another hospital, but currently be measured with cost per case-mix adjusted separation at an aggregate level as less efficient. This is illustrated with a simple hypothetical example in Table 2.4.2 for two hospitals (A, B) producing two DRGs (DRG1, DRG2) each with a case-mix weight of one (representing an average expected cost per admission of $7000).

Table 2.4.2: The Fox aggregation paradox with cost per case-mix adjusted separation

	Hospital A	*Hospital B*	*Industry average*
DRG 1			
Cost per admission	$7000	$8000	$7000
Admissions	900	100	(DRG weight=1)
DRG 2			
Cost per admission	$5000	$6000	$7000
Admissions	100	900	(DRG weight=1)
Total			
Cost	$680,000	$620,000	
Case-mix adjusted admissions	1000	1000	
Cost / case-mix adjusted admission	$6800	$6200	

Hospital A has a lower cost per admission in DRG1 and DRG2, but is measured with cost per case-mix adjusted separation as less efficient at an aggregate level with a $600 ($6800 vs. 6200) higher cost per case-mix adjusted separation. This paradoxical result can occur where cost shares by clinical activity (DRG), assumed largely exogenously determined by patients presenting for care, favour the comparative advantage of hospitals less efficient in each activity. The paradox also implies that two public hospitals with identical economic efficiency in each inpatient activity are likely to be measured as differently efficient at an aggregate level, due to differences in exogenously determined cost shares (patients presenting for care) by activity.

2.4.2 Is the aggregation paradox problematic for public hospitals?

For the Fox (1999) aggregation paradox to exist, requires only that firms are multi-product firms, while for it to be problematic requires that the share of effort devoted to each activity is determined exogenously. Assessing public hospitals against the first of these requirements, Evans [1984:191] notes that:

> "..if seen as firms they (public hospitals) are multi-product firms producing some combination of inpatient care, outpatient services, education, research and community services.".

Even within inpatient care, however, there is a multiplicity of outputs. These can be characterised as: a variety of physically dissimilar services (activities classified by

major diagnostic category (MDC)) and complexity (DRG activities within a MDC) and outputs arising from the requirements of individual patients (according to severity, co-morbidities or risk factors).

Even within inpatient services, public hospitals are therefore multi-product firms at least down to a DRG level. Given that public hospitals are multi-product firms, the paradox is problematic if public hospital cost shares for activities (DRGs) are seen as exogenously determined. A strong argument in favour of public hospital output shares (between DRGs) being determined exogenously (outside their control) is that public hospitals have a primary duty to treat people who present for care. Against this argument is the ability of hospitals to specialise, or not specialise, in some MDCs or DRGs, and variability in clinical practice in terms of the ability, at the margin, to assign patients to different (particularly adjacent) DRGs, or to refer patients to other hospitals or other parts of the health system.

The existence of waiting lists may also provide evidence that the mix of DRGs is partly endogenously determined, at least for some subset of DRGs or MDCs. Waiting lists are, however, by nature more observable in elective procedures, which in Australia constitute those DRGs with higher levels of private hospital activity. Consequently, in Australia, these activities are arguably of less importance in public hospital performance measures unless they are being compared with private hospitals. Despite exceptions at the margin, the mix of core inpatient activities for public hospitals is argued as exogenously determined by the mix of patients presenting for treatment. Table 2.4.3 describes characteristics of hospitals against requirements for the Fox (1999) aggregation paradox to exist, and be problematic, in performance measurement at different levels of analysis (DRG, major disease classification (MDC), inpatient care and all hospital output).

Table 2.4.3: Paradox conditions by level of analysis: public hospitals in Australia

Level of analysis	Paradox applies: Multi-product firm	Paradox problematic: Exogenously determined cost shares across activities
DRG	Homogeneous output by activity but not necessarily patient's risk factors or severity	Not applicable if DRG seen as lowest level of output. Service attributes and quality of care endogenously determined.
MDC	Activities within MDCs similar in nature, but not complexity.	Mainly exogenous but arguably partly endogenous. Limited movement between DRGs within MDCs (depends on clinical practice).
Inpatient care	Activities heterogenous by nature, complexity and patient risk factors.	Largely exogenous for public hospitals as patients cannot easily move between MDCs, but hospitals may not have certain MDCs and can influence DRGs within MDCs to a limited degree.
Hospital	Outputs intractably Heterogeneous	Mix of inpatients and outpatients largely determined exogenously

To avoid the paradox requires that activities are homogeneous. Activities are relatively homogeneous only at a DRG level, but even at this level ideally would require adjustment for patient risk factors, and should consider quality of care, as argued in section 2.2.

The paradox is problematic where cost shares between activities are exogenously determined. Cost shares between DRGs are largely exogenously determined for public hospitals at an aggregate inpatient level. There may be some MDCs for which treatment mode and type are discretionary, but for most MDCs the mix will still be largely determined by patients presenting for care. In general, at a hospital level the case-mix of public hospitals is out of their control. The characteristics of public hospitals, therefore, suggest the Fox (1999) aggregation paradox is problematic in efficiency measurement at levels above individual diagnostic related groups.

The aggregate efficiency measure of cost per case-mix adjusted separation, while weighting separations by an expected average relative cost, does not adjust for exogenously determined differences in cost shares across activities when calculated with costs and case-mix adjusted separations added up at an aggregate level. Hence the Fox (1999) aggregation paradox is faced with the current measurement of cost per case-mix adjusted separation measured at an aggregate level across hospital inpatient activities (DRGs). Therefore, if cost shares by inpatient activity are considered exogenously determined across hospitals, the paradox is problematic for aggregate economic efficiency measures, such as cost per case-mix adjusted separation.

In public hospitals in Australia these cost shares, can be seen as largely determined by patients presenting for care, and hence exogenously determined. However, for private hospitals in Australia cost shares by activity are under the hospital's control in the absence of a service obligation. Private hospitals can choose the cost shares across activities and which elective procedures to provide. While the paradox still occurs with private hospitals, it is not problematic as it reflects allocative efficiency in the choice of mix of clinical activities (DRGs). Such private hospitals can select an activity mix to maximise allocative efficiency across activities and hence aggregate economic efficiency.

2.5 Combined implications of performance-efficiency and aggregation paradoxes

In Australian public hospitals, if performance is measured with an aggregate activity-based economic efficiency measure, such as cost per case-mix adjusted separation, both paradoxes are faced.

In any bilateral comparison, a hospital with lower net benefit and/or higher cost per admission in each activity, can be measured as more efficient with cost per case-mix adjusted separation measure at a hospital level. In measuring relative performance, and particularly in identification of peers or benchmarks, cost per case-mix adjusted separation therefore clearly has the potential to be misleading and provide perverse incentives.

In practice, the Fox (1999) aggregation paradox, which is problematic due to the exogenous nature of cost shares by activity, may be considered less serious than the performance-efficiency paradox, which actively creates economic incentives for lower quality of care. However, the seemingly more benign nature of the aggregation paradox assumes that public hospitals treat patients presenting for care, and as a consequence, are inappropriately rewarded or punished for factors outside their control.

If hospitals are aware that they are punished if their largely exogenously determined patient mix does not match their comparative advantages in service delivery, economic incentives are created for public hospitals to avoid service obligations. That is, to not treat patients presenting in treatment activities where they lack a comparative advantage in cost per admission. These incentives might have some merit at a system level if referral based on comparative advantage in cost per admission reflected net benefit maximising behaviour. However, there is a lack of economic accountability for quality of care with current hospital case-mix performance measures and funding. Hence a comparative advantage in lower cost per admission for any given clinical activity (DRG) can reflect a lower quality of care and net benefit per admission.

Public hospitals are encouraged to focus on activities where they have a comparative advantage in cost per admission, but not necessarily a comparative advantage in net benefit per admission. Such referral practices may be seen as particularly problematic where lower cost per admission for any given clinical activity is attributable to lower quality of care, and can even reflect the worst performing hospital (lowest net benefit).

Incentives for lower quality of care created by the performance efficiency paradox can, therefore, be seen as a catalyst for turning the Fox (1999) aggregation paradox from benign to malignant. Public hospitals are provided with economic incentives to reduce access of patients to activities with high quality care but higher cost per admission, even though this may represent care with a comparative advantage in net benefit. Acting on such incentives also shirks requirements to treat patients presenting, and consequently acts against the nature of public hospitals as community institutions. At a health system level, such shirking can in turn lead to increased adverse health outcomes associated with delay from reduced access to, and lack of continuity, of care. For

patients it can also lead to greater opportunity costs of time and travel and reduced access to high quality care, or even care in general.

2.6 Summary

In chapter 2, the current performance measure of cost per case-mix adjusted admission, while partially overcoming cream-skimming, to the extent that patients within DRGs are homogenous, has been shown to create incentives for reduction in quality of care and cost-shifting. Economic efficiency measures such as cost per case-mix adjusted admission, by not incorporating quality or effects of care, face a performance-efficiency paradox. Hospitals measured as economically efficient with the lowest cost per case-mix adjusted separation, if attributable to lower quality of care, can be the worst performing, with lowest net benefit per admission.

The Fox (1999) aggregation paradox has also been demonstrated as applicable in comparing relative hospital performance with the aggregate economic efficiency measure of cost per case-mix adjusted separation. In implicitly aggregating efficiency for each clinical activity (DRG) by exogenously determined cost shares, public hospitals are rewarded or punished for factors outside their control. This allows paradoxical ordering of performance where peers can be identified who are less efficient in each clinical activity than other hospital/s. For public hospitals where cost shares by activity can be seen as largely exogenously determined by patients presenting for care, this has been illustrated as problematic.

The combination of the performance-efficiency and Fox (1999) aggregation paradoxes in using aggregate economic efficiency measures of performance, such as cost per case-mix adjusted separation and case-mix funding, have been suggested as particularly problematic. Economic incentives are not only created to shirk quality in care for each clinical activity, but also to shirk service obligations to treat patients presenting for care. In being economically accountable for cost, but not quality, of care, hospitals are encouraged to focus on activities with comparative advantage in cost per admission, but not necessarily net benefit per admission. Conversely, incentives are also provided to shirk service obligations in reducing access of patients to activities where hospitals have quality of care above a level minimising cost per admission. In measuring

performance with case-mix funding, ignoring quality and the exogenous nature of cost shares by activity, patients face both reduced access and lower quality care, without any guarantee of lower cost in treating these patient populations in the health system over time. In chapter 3, analysis at a clinical activity (DRG) level is identified as providing the ability to overcome both the performance-efficiency and Fox (1999) aggregation paradoxes.

Chapter 3: Level of aggregation for analysis

3.1 Overview

In this chapter, public hospital efficiency originating at a clinical activity (DRG), rather than aggregated hospital level, is suggested as required to overcome a performance-efficiency and aggregation paradox (Fox, 1999).

An individual clinical activity, diagnostic related group (DRG), level of analysis allows non-paradoxical aggregation and quality of care measurement, without confounding present at an aggregate level. A DRG level also allows adjustment for patient risk factors within-DRG, and linkage to post-care effects, in overcoming incentives for cream-skimming and cost-shifting respectively. Performance measurement, including quality at a DRG level, is identified as feasible in Australia using currently collected data for admissions, costs and disutility event rates as indicators of quality of care. Verifiable disutility bearing events currently collected in Australia include mortality, return to theatre, post surgical complication and other morbidity data, iatrogenic event e-code data and readmission to hospital. The potential also exists to link to post-separation events in other settings in Australia, with linkage to data sources such as death, mental health and cancer registries, nursing home care, and Health Insurance Commission (HIC) data for pharmaceutical and non-ambulatory care use and cost.

3.2 Clinical level performance analysis: overcoming paradoxes and creating appropriate incentives

3.2.1 Overcoming the performance-efficiency paradox

A solution to the performance-efficiency paradox requires the ability to include outcomes or quality of care indicators. However, the appropriate type, and relative importance, of clinical quality of care indicators differs by clinical activity (DRG). Depending on clinical activity mortality, specific morbidities or complications of surgery, adverse drug reactions or events post-care such as readmission by specific DRG, may be appropriate indicators of clinical quality of care.

At a clinical activity (DRG) level, decision analytic methods can be used to flexibly identify effects within and beyond hospital separation, and patient risk factors can be identified, and adjusted for. In comparison, at an aggregate level, choice and

measurement of global indicators of quality (disutility event rates) are confounded across activities by differences in patient risk, by activity, and degree of attribution to quality of care.

Relative differences in hospital performance are affected at an aggregate level by exogenously determined shares across activities with different baseline risks. If disutility rates were included at an aggregate level, hospitals with a mix of activities with greater base risk of harms would be disadvantaged, unless standardised for risk by case-mix between activities. However, while case-mix adjustment could be employed to adjust for differences in risk across activities, it cannot allow for differences in attribution to quality of care or appropriateness of indicators, and does not adjust for patient factors within-DRG. Additionally, given heterogeneity of indicators, any set of quality indicators identified at an aggregate level will only be a partial set of indicators. Using quality indicators in performance measurement or funding at an aggregate level, therefore, provides scope for finessing by providers in reducing quality in aspects not measured. Even with measured aspects of quality, providers can comply with a global indicator in activities where the indicator does not reflect quality of care, or is less meaningful.

In performance measurement at an aggregate level, meaningless input-output ratios can also arise. For example, the meaningless ratio of outpatient days per bed was implicitly allowed for in aggregate studies of hospital efficiency by Hogan and Wrobleweski (1993), Holvad and Hougaard (1993), Mangnussen (1996), Ferrier and Valadmanis (1996) and Mobley and Magnussen (1998).

3.2.2 Allowing for cream-skimming and cost-shifting

A DRG level of relative performance measurement increases homogeneity of patient admissions compared with more aggregate levels of analysis. At a DRG level of analysis adjustment to concentrate on exogenously determined risk factors within DRG, such as age, sex and co-morbidities, rather than case-mix adjustment in aggregating across DRGs. Consequently, at a clinical activity (DRG) level adjustment for patient risk factors at point of admission becomes feasible in performance measurement (addressed in section 5.7.1) or funding (addressed in section 6.8). Where patient mix within-DRG can be endogenously determined by hospitals adjustment for these risk

factors allows economic incentives for cream-skimming to be overcome. At a DRG level, decision analytic methods, can be used to comprehensively identify effects, resource use and costs of care, including those beyond-separation to prevent incentives for cost and event shifting.

As with event rates and costs within-admission rates of events and costs beyond discharge should be adjusted for patient risk factors at admission, , but also degree of attribution to care. To allow for increasing role of external effects over time, degree of attribution could be made dependent on time (i.e. a dampening function) as well as scope of events post-admission (for example specific readmission DRGs could be identified using decision analytic methods).

3.2.3 Avoiding the Fox (1999) aggregation paradox

The Fox (1999) aggregation paradox is preventable by considering relative efficiency for each clinical activity individually, as there are then no cost shares to differ between hospitals (as outlined in table 2.4.3, section 2.4.1). An additional advantage of considering performance at a clinical activity level is that relative efficiency across individual hospital activities can be identified. A clinical activity level therefore reveals inefficiencies in individual activities, hidden by aggregation, allowing appropriate peers and benchmarks for each activity to be identified. Policy makers may, however, also want to consider more aggregated measures of hospital performance, for example at a major disease classification (MDC) or inpatient hospital level.

3.2.4 Aggregating efficiency to avoid the Fox paradox

For multi-product firms, if a summary measure of efficiency is required, then Aczel (1990) demonstrated that a geometric mean aggregating lower level (DRG level) efficiency scores is to be preferred to competing means. Fox (1999:175) noted that a symmetric (unweighted) geometric mean avoids the aggregation paradox. An unweighted geometric mean provides a summary measure of efficiency E for any hospital from efficiency in each of their n activities E_i $(i = 1..n)$ as:

$$E = \prod_{i=1}^{n} E_i^{1/n} \qquad (3.2.1)$$

However, this approach does not incorporate the relative importance of activities, giving equal weight to diagnostic related group activities, whether trivial or important in their cost share for the individual hospital or for the health system generally. Conversely, the use of individual hospital cost-shares to weight outputs, although incorporating the relative importance of different outputs in a hospital's measure of efficiency, as discussed in section 2.4, faces the Fox (1999) aggregation paradox.

A solution I propose which both avoids the aggregation paradox and maintains the relative importance of activities, is the use of standard industry cost shares by activity to weight individual DRG level efficiency scores for a hospital, into a summary geometric mean. For each hospital *j (j=1…m)* with outputs *i(i=1...n),* aggregate economic efficiency E_j can be estimated from efficiency in each activity (output) as:

$$E_j = \prod_{i=1}^{n} E_{ji}^{S_i} \qquad (3.2.2)$$

where E_j is overall economic efficiency for hospital *j* (*j*=1... m), E_{ji} is hospital *j's* efficiency in activity *i* and S_i is the standard (industry) cost share in activity *i.* Such a solution is analogous to the use of age standardisation to overcome the Simpson (1951) paradox The aggregation method proposed in equation (3.2.2) compares public hospitals with different case-mixes as though they were the same, while maintaining the relative importance of activities at an industry level. The practical problem of individual hospitals not having DRG separations for any given DRG can be simply adjusted for by re-weighting the standardized cost shares to standardised industry shares without those DRGs. Therefore, using standardised industry cost shares, policy makers can be informed of hospital performance and appropriate peers and benchmarks at any level of aggregation across inpatient services, without facing problems of the Fox (1999) aggregation paradox. The same method can also be applied to aggregate efficiency scores for non-inpatient, or any other hospital activity, in compared hospitals.

3.2.5 Non-paradoxical aggregation of cost per case-mix adjusted separation

At present, cost per case-mix adjusted separation is implicitly measured for each hospital by adding costs of all diagnostic related group admissions (at a hospital level) and dividing by case-mix adjusted separations. As shown in table 2.4.2 this aggregation method faces the Fox (1999) aggregation paradox. However, this paradox can be avoided and the relative industry importance of DRGs preserved if cost per case-mix adjusted separation were measured for each hospital for each DRG, then aggregation undertaken using fixed industry cost shares. Two hospitals with the same cost per admission in each DRG would then have the same aggregated cost per DRG adjusted separation and, in any bilateral comparison, a hospital with lower cost per separation for each DRG, would have lower cost per DRG adjusted separation. Never the less, the performance-efficiency paradox would remain problematic.

3.3 Feasibility of DRG level performance analysis

Having established the ability of a DRG level performance analysis to overcome problems at an aggregate level, a natural empirical question is whether a DRG level of analysis is feasible. What data for public hospitals are available at a DRG level to allow efficiency measurement including quality of care?

3.3.1 Australian DRG level cost and activity data: The National Hospital Cost Data Collection

In Australia, the annual National Hospital Cost Data Collection (NHCDC) used in estimating DRG cost weights can be used to identify, for each hospital in the NHCDC sample, DRG patient separations and their associated costs by sixteen sub-components. Efficiency analysis can therefore be undertaken at any level down to that of the individual DRG for hospitals within the NHCDC sample. In 1998-99 this sample included 150 public hospitals in Australia, representing 64% of separations, with 72% of large teaching hospitals participating (DHAC, 2000).

Cost by DRG are estimated in this sample for the sixteen components: ward medical, nursing medical, allied health, imaging, pathology, pharmacy, critical care, operating rooms, emergency department, supplies, hotel, on-costs, depreciation, prosthetics, special procedure suites and other. In determining DRG level costs, there is potential

for differences in accounting practices, as Street (2003) stressed in his UK analysis of hospital performance measurement. In Australia, conformity has increased with annual cycles of collecting NHCDC data for diagnostic related group case-mix weights since 1992 (Australian Refined Diagnostic Related Group (AR-DRG) case-mix weights since 1997). Nevertheless NHCDC data in Australia is allocated mainly[1] in a top down method from a hospital level to DRG level, providing potential for gaming in allocating costs across clinical activities (DRGs). Ideally, the coverage of all DRGs would provide a natural way of preventing finessing in allocation of cost data to individual activity by hospitals. Accuracy of costs allocated to individual DRG may be potentially biased where costs are allocated top down, particularly in allocating joint costs between activities.

In addressing accuracy in allocating joint costs, a major disease classification (MDC) level of analysis offers the potential for reduced bias in top down allocation than a DRG level. More than 660 DRGs condense into 23 MDCs, and hence the scope for misallocation is smaller. As Harris (1977) noted, many costs are specific to MDC which are arranged along lines of clinical specialty:

> "the resource transfers between medical staff and ancillary departments go predominantly in certain directions. The cardiac catheterisation labs are used primarily be cardiologists. The operating rooms are used primarily by surgeons. Special orthopaedic appliances are ordered primarily by orthopaedic surgeons. Brain scans and brain angiograms are ordered primarily by neurologists."
> Harris (1977:481)

A MDC level still retains some advantages of a DRG level of analysis for including quality, while avoiding potential biases in allocation of joint costs from MDC to DRG level. However, at an MDC level, case-mix weights are still likely to be required in allowing homogenous measurement of output to aggregate the approximately 30 DRGs within each MDC. As case-mix weights are derived from the NHCDC sample they will still reflect any systematic biases in cost allocated to DRG across the sample. In this respect, using NHCDC cost data at a DRG level to analyse performance, makes no

[1] A proportion of Victorian hospitals in the NHCDC sample provide data using 'bottom up' micro costing methods.

more assumptions in attribution of costs to individual DRG's than are already implicitly made in calculating case-mix weights using NHCDC data.

3.4 Quality of care indicators at a DRG level

In Australia, the Australian Council on Healthcare Standards (ACHS) routinely collects data from more than 500 acute care public hospitals on approximately 185 clinical indicators of outcome and processes, across a wide range of specialties (Collopy, 2000; Boyce et al., 1997; Fahey and Gibberd, 1995). Since 1993, Australian hospitals have been required to provide this data in accreditation or re-accreditation (Howley and Gibberd, 2003). Potential verifiable quality indicators, provided by this accreditation data, include return to theatre, drug reactions and adverse event causes of mortality and morbidity. This data set also includes many other variables specific to DRG. For example, in gynaecology and obstetrics, indicators include: Apgar scores; primary caesarean section for failure to progress or foetal distress; hysterectomy in women below the age of 35, and; urinary tract infection during a gynaecological procedure.

ICD-10 codes of iatrogenic harm (e-codes), from the Australian Institute of Health and Welfare's (AIHW) national morbidity and mortality database are an alternative source of routinely collected and verifiable quality of care indicators (Hargreaves (2001)). Time dependent rates of quality indicators, such as readmission to hospital by DRG, utilisation of pharmaceuticals and non-ambulatory care from linkage to HIC data, or rate post hospitalisation mortality in linking to death registry data, could also be constructed at a DRG level. Data linkage has already been used by Gregan and Bruce (1997) to estimate readmission rates in a DEA study of relative Victorian hospital performance, explored in detail in chapter 4,.

3.4.1 Using quality of care indicators at a clinical activity level

To allow performance to reflect the objective functions of public hospitals at a DRG level requires inclusion of health effects or related quality of care indicators. However, caution needs to be adopted, as health effects in patient populations are not only determined by health services in hospitals, depending on the patient and their risk factors in the process of care, and can take time to evolve. Depending on DRG, capturing health effects attributable to hospital care may, therefore, require linkage to

post-care effects as well as verifiable aspects of effects within-care (mortality, drug reactions, iatrogenic harms), and adjustment for patient risk factors.

In reporting on the AIHW adverse event data, Hargreaves (2001:2) noted that:
"by use of the words 'unintended' or 'unnecessary' the Australian Patient Safety Foundation' definition of adverse events includes only harm or suffering that is not unavoidable. Thus conditions considered as normal, or expected consequences of treatment are not included." As an example of this, adverse events such as ICD-9 external cause codes E870-E876 (misadventures to patients during surgical and medical care) or E878-879 (surgical and medical procedures as the cause of abnormal reaction of patient, or later complication without mention at the time of procedure) could be used without problems. However, external cause codes such as E850-858 (accidental poising by drugs medications and biological cause) would include heroin overdoses, not defined as adverse events attributable to care under these definitions. The conservative nature of the patient safety foundation definition, in excluding factors not directly attributable to hospital care, ensures defined events are appropriate for consideration in performance analysis of hospitals. However, this definition may be too restrictive to allow a comprehensive consideration of health effects of quality of care with effects that may not be directly attributable. In general comparing rates of these adverse events require adjustment for patient risk factors at admission, and effects beyond-separation given discretion with respect to timing of separation.

3.4.2 An example of including disutility events as quality indicators at a clinical activity level

A recent study by Simpson et al. (2003) compared quality of care in pre-term babies using rates of severe intraventricular haemorrhage (IVH) adjusted for patient characteristics at a clinical activity level. Rates were adjusted for prognostic factors (within DRG case-mix) using logistic regression on five significant predictor variables: gestation age; 1 minute Apgar score; antenatal corticosteroids; transfer after birth and; gender. The observed relative to the expected event rate were also adjusted for sampling variation using Bayesian shrinkage estimation methods (explored in detail in chapter 9).

This study by Simpson et al. (2003) provides an example of the ability to appropriately identify a quality of care indicator at a DRG level, which can then be adjusted for patient level risk factors. At a clinical activity level rates of disutility events such as mortality, morbidity, readmission and iatrogenic events identified to reflect quality of care in a meaningful way, and risk factor adjustment can focus on patient characteristics within DRG.

3.5 Summary

A DRG level of performance analysis has been identified as enabling more appropriate relative performance than an aggregate level in allowing:

(1) flexible inclusion of quality of care or health effects (such as mortality, mortality or other iatrogenic events, readmissions) without confounding across activities;
(2) costs and event rates to be standardised on prognostic factors, preventing incentives for cream-skimming;
(3) identification of attributable events for data linkage using decision analytic methods, preventing incentives for cost and event shifting;
(4) inefficiency hidden by aggregation to be revealed and;
(5) the Fox (1999) aggregation paradox to be avoided, with DRG level economic efficiency able to be non-paradoxically aggregated using industry cost shares.

DRG level performance is feasible using NHCDC cost and admission data (CDHS 2000) and verifiable disutility event rates such as mortality, surgical complications and other morbidity, adverse drug reactions, e-code data, return to theatre and readmission (AIHW, ACHS, State Health Authorities) as indicators of quality of care.

Chapter 4: Performance measurement with quality, specifying disutility events as outputs

4.1 Overview

To measure relative performance across public hospitals allowing for effects of quality of services, Data Envelopment Analysis (DEA) is suggested to have advantages over alternative efficiency measurement methods. DEA can handle multiple inputs and outputs without prices or parametric assumptions, and input-orientated economic efficiency can be decomposed to inform policy and decision makers of, allocative, technical and scale inefficiency. Using the scale invariance property of radially contracted distance functions, DEA is shown to also allow technical efficiency to be estimated when only cost by input factor is available, under the assumption that hospitals face the same prices for factor inputs.

Attempts to specify quality, represented by disutility event rates such as mortality, morbidity and readmission as an output with DEA, are, however, demonstrated to encounter problems in flexibly representing an appropriate economic objective. Output specifications of disutility events examined include:

(1) inverted rates of disutility events;

(2) admission without disutility events and;

(3) the hyperbolic approach of Färe, Grosskopf, Lovell and Parsuka (1989), with equi-proportional expansion of desirable outputs, and contraction of undesirable outputs.

Specifying inverted rates of disutility events as an output is illustrated to face size-biasing, unless normalised by level of activity (admissions). Even with normalising, non-linearity and dimensionality problems remain in trading off quality and quantity of care.

The absence of relative prices for outputs of admissions, and inverted disutility event rates precludes considering output-orientated economic efficiency. Consequently, problems with technical efficiency measures of performance under this specification of hospitals with high disutility rates being able to compete on cost per admission, and hospitals with high cost per admission competing on quality alone, can only be

mitigated to a limited degree in using approaches such as assurance regions (Thompson 1986, 1990, 1992) in restricting output weights. In general, an inability to provide an appropriate trade-off between quality represented by inverted rates, and quantity of services, creates perverse incentives for high and low quality of care.

An output specification of admissions without disutility events ensures a trade-off for all hospitals between quality and quantity of care. However, the value of admissions with disutility events is restricted to 0. This implicit value of 0 may be appropriate where disutility events are readmission to the same DRG. However, in general, this specification does not allow a flexible trade-off for other disutility events where the value of admission with disutility bearing events may be negative, or positive, depending on disutility event severity. For a single type of disutility bearing health effect, this specification corresponds to an economic objective of minimising average cost-effectiveness (the ratio of costs relative to effects). Average cost per unit of health effect is, however, problematic as a relative performance measure for hospitals in failing to allow for the incremental and non-tradable nature of health effects in given patient populations. The value of incremental effects relative to incremental costs (Drummond et al., 1987; Drummond et al., 1997), implicit in maximising net benefit, is suggested as preferable in representing relative performance of hospital providers, as it is in health technology assessment comparing technologies.

The hyperbolic method of Färe, Grosskopf, Lovell and Parsuka (1989), with equi-proportional radial contraction of undesirable outputs and expansion of desirable outputs, is demonstrated to face problems in translating to a setting where disutility events represent effects or quality of care. Activity and disutility events are not meaningfully separated as desirable and undesirable outputs in the same way that marketable desirable outputs (such as electricity generation), and undesirable outputs, (such as pollution) are. Assuming hospital admissions *per se* are desirable is problematic, in implicitly ascribing value without considering health effects and effectively ignoring the derived nature of demand for health care. Where rates of undesirable outputs represent quality of care, regions of the hyperbolic frontier where admissions decrease, and disutility event rates increase, make interpretation of technical efficiency, as a performance measure, problematic. In the absence of economic efficiency measures, technical efficiency improving where disutility event rates and

costs per admission increase cannot be subsumed into allocative efficiency, representing perverse choice of quality of care. Congestion efficiency, estimated under the hyperbolic method as the residual of technical efficiency under strong (costless) and weak disposability of disutility events, is also shown to not have a meaningful interpretation where disutility events represent quality of services. In hospitals, estimation of congestion efficiency under the hyperbolic method perversely represents the maximum increase in admissions possible for given resources, without constraints on disutility event rates. Finally, the method of Färe, Grosskopf, Lovell and Yaisawarang (1993), used to estimate a monetary shadow price for undesirable outputs, cannot be employed in the absence of a market price for admissions *per se*.

Each attempt to measure hospital performance including disutility event rates (mortality, morbidity, readmission) specified as outputs are, consequently, suggested to be unable to provide an appropriate underlying economic objective to flexibly trade-off cost and value of quality (health effects) of care. Critical analysis of output specification problems in relative performance measurement does, however, suggest an objective of net benefit maximisation and input specification of disutility event rates in representing health related quality of care.

4.2 Choice of efficiency measurement method

In chapter 3 a clinical diagnostic related group level was identified as preferred to aggregate levels, in identifying relative performance measurement allowing for health related quality of care, represented by disutility-bearing event rates such as mortality, morbidity and readmission. However, questions remains as to the method of efficiency measurement and specification of disutility events in efficiency measurement to reflect an underlying appropriate objective function.

To measure relative economic performance of public hospitals with available data, we ideally require a method that can handle multiple inputs and outputs in the absence of relative prices for outputs, and the flexibility to allow for disutility events to represent effects of quality. Candidate methods for performance measurement include: data envelopment analysis (DEA), stochastic frontier analysis (SFA), least squares (LS)

econometric production models and total factor productivity (TFP) indices (e.g. Tornquvist/Fischer).

Assessing these alternative methods against the requirements, DEA appears to best fit the criteria. DEA allows multiple inputs and outputs, does not require prices or pre-specified weights (unlike TFP) or the *a priori* specification of functional form (as SFA or LS do). DEA can also be used to identify peers and targets (unlike TFP), can decompose technical efficiency under constant returns to scale (CRS) into technical efficiency under variable returns to scale (VRS) and scale efficiency, without behavioural assumptions (unlike LS or SFA). Cost and residual allocative efficiency can also be estimated with behavioural assumptions, such as cost minimisation (as SFA requires).[1]

Appendix 4.1 outlines the DEA method and formulations under constant (Charnes, Cooper and Rhodes, 1978), variable (Banker and Charnes, 1984), and non-increasing (Coelli, Rao and Batesse 1998:151; Färe, Grosskopf and Lovell, 1994:50) returns to scale, from an input-orientation. DEA also has the ability to estimate technical efficiency with cost data for factor inputs, using the unit invariance property of Farrell (1957) radial distance functions as outlined in appendix 4.2. However, this requires the assumption that producers (decision making units) face the same set of prices for factor inputs.

The choice of DEA over SFA, in modelling hospital performance, reflects the advice of Hollingsworth, Dawson and Maniadakis (1999), who, in reviewing frontier studies of efficiency in health care, suggested a preference on balance for DEA due to flexibility and lack of *a priori* assumptions. The advantages of these flexibilities become even more apparent in allowing for quality of care.

DEA has been chosen as a preferred method, due to its flexibility and absence of *a priori* assumptions in representing technology allowing for quality. However, it should be noted that DEA is generally non-parametric, and that this flexibility comes at the

[1] Further advantages of DEA over SFA in allowing for quality of care in performance measurement are explored in considering the use of SFA in application of the correspondence theorem in section 8.4.3 of chapter 8.

price of not accounting for uncertainty. In comparison with stochastic frontier analysis (SFA), non-parametric DEA does not, therefore, allow conventional statistical testing of hypotheses (Coelli, Rao and Batesse 1998:243-246). However, in modelling uncertainty, SFA faces the problem of a largely non-testable *a priori* assumption in determining an empirical distinction between error terms and inefficiency, as well as a potential for misspecification of functional form for technology, considered at length in section 8.4. DEA, while non-parametric, still requires enough observations to allow observed 'best practice' of the frontier to represent technology. DEA does not, therefore, completely escape issues of parsimony in choice of variables relative to the number of observations.

While not requiring a functional form, the appropriateness of the objective function for hospitals, implicit in the relationship between modelled inputs and outputs, should be the driving factor in choice and specification of variables for DEA. The objective functions that performance measurement and funding mechanisms implicitly represent, are particularly important to consider in hospitals, given the intermediate nature of activity and the absence of transaction conditions for a market.

The multiple nature of inputs (particularly specialist forms of labour, capital and disposable inputs), throughput measures (occupancy rates, length of stay) and the heterogenous activities as outputs further reinforces this importance of the underlying objective function the specification represents. As Eckermann (1994:169) noted:

> " ..hospital efficiency requires careful conceptual consideration before assessing how current measures of hospital activity can be used …There is a temptation to use any available data to assess efficiency without due recognition of the need to determine a sound theoretical basis on which to do so."

A clinical activity (DRG) level of analysis has been identified as the appropriate level to include quality with disutility event rates, and DEA is suggested as the most appropriate method to allow hospital performance measurement, including quality. In attempting to integrate quality of care into performance measurement at a clinical activity level, the main question this thesis addresses is how to include quality of care, represented by effects of care such as mortality, morbidity and readmission, to reflect an appropriate objective function. In the remainder of this chapter, attempts to provide

an appropriate underlying objective function in efficiency measurement, allowing for quality of care with output specifications of disutility events, are examined and critically analysed. While these attempts are shown to be fruitless, they point to an input specification, which, in chapter 5, is identified as able to provide relative efficiency measurement consistent with an underlying objective of net benefit maximisation.

4.3 Specifying quality in efficiency measurement

In modelling hospital performance to reflect an appropriate objective, Newhouse (1970:66) proposed a trade-off between quality and quantity of care for given resource use with 'constrained quality-quantity maximization'. While this trade-off implicitly suggests an output specification for quality, how quality was to be measured or incorporated in specifying performance as an output was, however, not addressed other than to:

> "assume quality to be represented by a vector of characteristics".
>
> (Newhouse, 1970:66).

Harris (1977) implicitly addresses what quality could be represented by, in suggesting a trade-off between the primary objective of medical staff of health maximisation representing quality of care, and cost minimisation as the primary objective of hospital administrators. In doing so, Harris infers that focusing on cost minimisation and ignoring health maximisation (quality of care) is problematic when he states:

> "..our current regulatory policy is almost exclusively directed at the supply side of the organisation. Unless we revise our definition of hospital to include the 'doctor' part of the firm the policy is doomed to failure." (Harris, 1977:467)

In measuring relative performance of hospitals, while consideration of effects of care is required to prevent perverse incentives for quality of care, health effects are not easily incorporated, in a meaningful way to an efficiency framework. Health, if considered as a variable, has characteristics of a stock variable, which health care may change incrementally (Grossman, 1972). Comparison of relative performance therefore needs to allow the ability to remove floor effects of health as a stock variable in any patient population. Health, if considered as a commodity, is also not a tradable commodity

but rather specific to patients treated. As McGuire, Henderson and Mooney (1988:32) state:

> "Health itself is not tradeable in the sense it cannot, strictly, be bought or sold in a market... health is not exchangeable."

Hence, in representing trade-offs between health effects and resource use (or costs of care) in performance measurement, it is the absolute incremental health effect of care traded off against incremental costs that is appropriate, rather than average cost-effectiveness.

Numerous studies of hospital performance have been undertaken using the frontier methods of DEA and SFA. These studies include those of Sherman (1984); Banker, Conrad and Strauss (1986); Grosskopf and Valdmanis (1987); Long (1990); Morey, Fine, Loree, Retlaff-Roberts and Tsubakitani (1992); Valdmanis (1992); Färe, Grosskopf, Lingdren and Roos (1993); Zuckermann, Hadley and Lezzoni (1994); Grosskopf, Margaritis and Valdmanis (1995); Thanasoulis, Boussofiane and Dyson (1995); Landon et al. (1996); Bruce and Gregan (1997); Mobley (1998); Puig-Junoy (1998); Webster, Kennedy and Johnson (1998) and; Street (2003). As Folland and Hoffler (2001:683) noted, frontier analysis of hospital inefficiency (initially using DEA and increasingly SFA) has become a minor industry. Hollingsworth, Dawson and Maniadakis (1999) identified almost 90 applications of DEA to hospital in Europe and the US alone.

The vast majority of these studies have been at the aggregate level of the hospital with activity based measure/s of output, and most have empirically ignored health related quality of care. A limitation, noted in the conclusions to many of these studies using intermediate activity-based outputs, is interpreting results of relative efficiency measurement in the absence of health effects or quality of care.

Despite the clear imperative to allow for effects, as well as the costs, of quality of care in assessing relative performance, very few studies of hospital performance have attempted to model these effects. Morey, Fine, Loree, Retlaff-Roberts and Tsubakitani (1992), Zuckermann, Hadley and Lezonni (1994), Gregan and Bruce (1997) and Puig-Junoy (1998) have attempted to model disutility events such as mortality, morbidity or

readmission, as effects of quality of care. With the exception of Morey et al. (1992), these studies have implicitly attempted to specify disutility events as outputs, and only the study of Puig-Junoy (1998) compared hospital performance at a clinical activity level. Zuckermann, Hadley and Lezzoni (1994), in estimating predicted, relative to actual, costs using SFA, included an indicator of whether hospitals' case-mix adjusted mortality rate was in the lowest or highest decile of case-mix adjusted mortality rate. Bruce and Gregan (1997) specified inverted readmission rates, in addition to admissions, as an output in DEA models of technical efficiency. Puig-Junoy (1998) specified days of survival and survival status as outputs in a DEA study of relative performance in neonatal care.

Outside of a hospital setting, Färe, Grosskopf, Lovell and Parsuka (1989) identified a hyperbolic method to allow comparison in equi-proportionally increasing desirable outputs (such as electricity generation) and reducing weakly disposable undesirable outputs (such as pollution). In this thesis each of these performance measurement methods for including disutility bearing effects of quality of care (such as mortality, morbidity and readmission) are critically assessed, against their ability to represent an appropriate underlying objective function, and provide appropriate economic incentives. In this chapter, the DEA output specifications of inverted disutility events rates of Bruce and Gregan (1997), admission without disutility event specifications as implicitly modelled by Puig-Junoy (1998), and the hyperbolic approach of Färe, Grosskopf, Lovell and Parsuka (1989), are considered. The implicit input-orientated DEA approach of Morey et al. (1992) is considered in section 5.8 and the SFA approach of Zuckermann, Hadley and Lezonni (1994) in section 8.4.2.

4.4 Specifying disutility events as outputs

4.4.1 Normalising inverted disutility-bearing event rates

Quality of care represented by disutility event rates can be considered a characteristic of admissions for a given hospital. As a general principle, where quality is specified, with disutility event rates (e.g. mortality, morbidity, readmission) as an output in addition to admissions, hospitals with the same disutility event rate should have the same output factor proportions between the specification of this rate as a quality variable and admissions. If this principle is violated, then comparison of hospitals can be seen as

biased by differences in size of compared hospitals. Quality, measured as an output with inverted rates of disutility bearing events (e.g. inverted mortality rates) in addition to admissions, violates this principle.

Output ratios between inverted disutility event rates and admissions are biased to favour smaller scale of production, in the absence of normalising by admissions. A specification including inverted rates of disutility-bearing events, in addition to admissions, as an output, efficiency measured for a hospital with a low number of admissions will improve relative to that for a hospital with the same disutility event rate (quality of care), but greater number of admissions. To illustrate, consider two hospitals, say A and B, with the same mortality rate (say 10%), but A has 20 admissions and B 100 admissions, as presented in table 4.4.1.

Table 4.4.1: Normalising inverted rates to avoid size-biasing

	Hospital A	*Hospital B*
Relative (to best practice) efficiency		
Mortality rate	10%	10%
Admissions (y1)	20	100
Inverted mortality rate (y2)	10	10
Ratio y2/y1	0.5	0.1
y3=y1×y2=admissions×inverted mortality rate	200	1000
Ratio y3/y1	10	10

If quality of care is assumed to be represented by mortality rate, then, for the same mortality rate (10% for hospital A and B in table 4.4.1), we would want the same relative output proportions between admissions and quality specified as an output. However, with quality specified as an inverted mortality rate, as the output quality measure is the same for each hospital with the same quality (10=1/0.1) regardless of size, quality per admission (output proportions) is measured as greater for the smaller hospital. Relative quality to admission is measured as 0.5 for hospital A and 0.1 for hospital B. The output proportion of quality to admissions is less for B, despite having the same mortality rate per admission. Hence in allowing for health related quality of care with disutility event rates this specification biases in favour of smaller hospitals, or

equivalently, against larger hospitals. Multiplying inverted mortality rates by admissions removes the potential for size-biasing by ensuring that output proportions are the same for the same disutility event rate regardless of admissions (10 for both hospitals A and B in table 4.4.1).

In general, if an inverted disutility-bearing event rate is specified to represent quality as an output, then for the same quality of care (disutility event rates), smaller hospitals are favoured in efficiency analysis relative to large hospitals. Smaller hospitals have greater proportional increase in outputs for the same disutility-bearing event rate, biasing relative efficiency estimates in favour of smaller hospitals, or equivalently against larger hospitals. To avoid size-biasing, inverted disutility event rates require normalising relative to other inputs and outputs. If inverted disutility event rates are to represent quality of care, they should be multiplied by admissions to be used in addition to admissions. Output ratios are then guaranteed to be the same for the same disutility event rate, regardless of hospital size. In this respect, inverted disutility event rates are no different from utility bearing event rates, where for example, average student test scores would need to be multiplied by number of students, if included as an output variable alongside students, in estimating relative teaching performance allowing for quality.

To illustrate the requirement to normalise in avoiding the effect of size-biasing in a DEA framework, consider where hospitals A and B in table 4.4.1 have the same cost per admission as well as mortality rate. Figure 4.1 depicts these two hospitals relative to an output-orientated frontier, under constant returns to scale, with quality specified in part (a) as inverted disutility event rates, and normalised inverted disutility event rates in part (b).

Figure 4.1(a): Size-biasing of efficiency with inverted disutility event rates

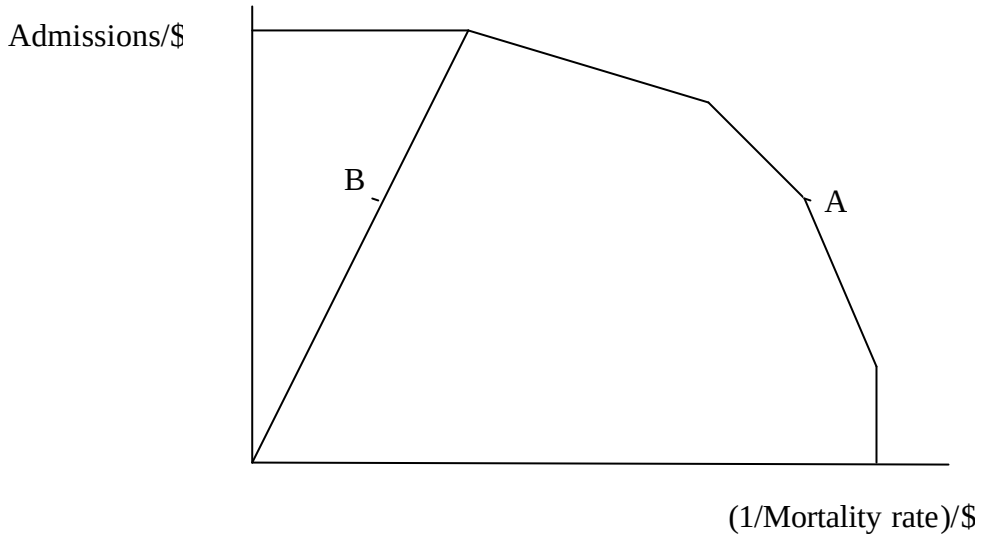
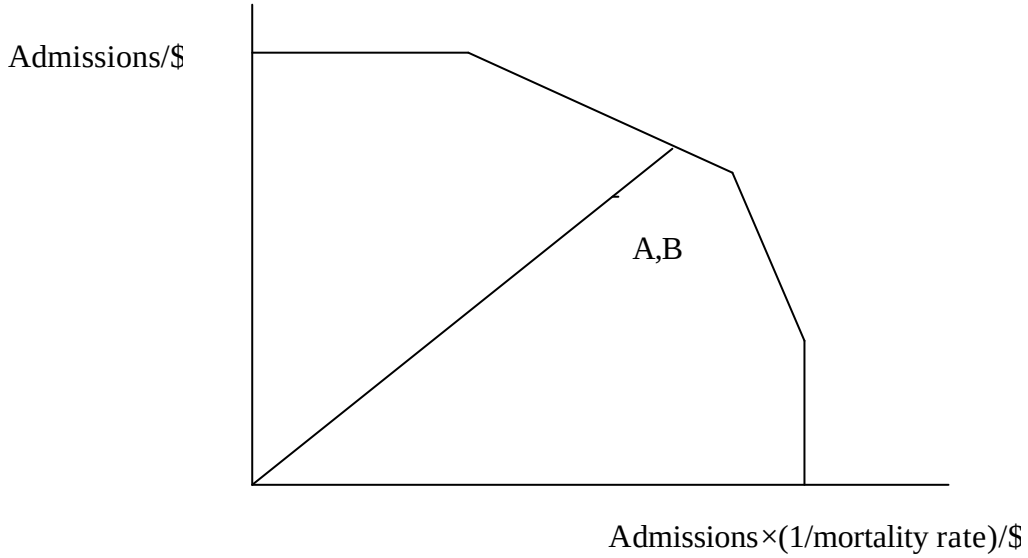


Figure 4.1(b) Overcoming size-biasing in normalising disutility event rates



While Hospital A and B have the same cost per admission for the same disutility event rate, it is only when inverted disutility events are normalised in multiplying by admission that they have the same efficiency score under constant returns to scale. Without normalising, despite having the same cost and quality per admission the estimated technical efficiency of the larger hospital is below that of the smaller hospital. In an output-orientation under constant returns to scale, the radial expansion to a convex frontier maximising admissions and inverted disutility event rates for given inputs, is greater for the larger hospital. Efficiency is, therefore, clearly size-biased using inverted disutility event rates in representing quality of care under constant returns to scale, but is corrected normalising on admissions.

The requirement for normalising inverted disutility bearing event rates to avoid size-biasing is important to establish. The only Australian study to attempt incorporation of quality of care effects in measuring hospital performance (Gregan and Bruce, 1997) used the inverted rate of readmission as an output, in addition to admissions. As this specification biases against larger hospitals, relative to smaller hospitals, in relative performance measurement including quality, empirical findings directly or indirectly related to size will reflect this bias. This would particularly be the case for the finding that:

> "Nearly all metropolitan/large country hospitals were relatively less efficient and therefore had small rural hospitals as their peers or benchmark partners." Gregan and Bruce (1997:63)

4.5 Comparing alternative output specifications of effects of care: Respiratory infection (DRG E62a) and mortality as a case example

To illustrate alternative output specifications, relative performance of 45 New South Wales public hospitals with 10 or greater admissions for the diagnostic related group of respiratory infection with complicating conditions (DRG e62a) are compared. Mortality was identified as the primary quality of care indicator for this DRG, with an average mortality rate of 22.4% across 45 hospitals.

To allow for the lack of a defined functional form to embody production technology, a DEA approach requires parsimony in choice of inputs and outputs, such that there are

enough decision making units (hospitals) to make the frontier of best practice a meaningful representation of technology in estimating technical efficiency (Coelli, Rao and Battese, 1998). Given a requirement for parsimony in constructing a best practice frontier for 45 hospitals, the 17 NHCDC subcategories of hospital costs were collapsed to 6 major sub-categories (% total expenditure): ward nursing (30%); ward medical (10%); pharmacy (9%); pathology (7%); critical care (8%) and other (37%).

4.5.1 Calculating technical, economic, allocative and scale efficiency

Input-orientated technical efficiency was estimated using Farrell (1957) distance functions with cost by factor input, under the assumption that all hospitals faced the same factor input prices, using the unit invariance property of radial contraction (see appendix 4.2).

Cost efficiency was estimated, assuming cost minimisation as an objective at any given quality of care. All calculations of technical and economic efficiency were undertaken using 'On Front version 2', following linear programming DEA formulations in appendix 4.1. Decompositions were implicitly undertaken for:

1. Allocative efficiency, calculated as a residual from the ratio of economic to technical efficiency under constant returns to scale (Charnes, Cooper and Rhodes, 1978);
2. Scale efficiency, calculated as a residual from the ratio of technical efficiency under constant returns to scale, and technical efficiency under variable returns to scale (Färe, Grosskopf and Logan, 1983).

4.5.2 Alternate DEA Specifications of quality of care as an output

Alternative output specifications were modelled for:

1. admissions alone;
2. admissions and inverted mortality rate;
3. admissions and inverted mortality rate multiplied by admissions and;
4. survivors alone.

The first output specification, equivalent to cost per admission for a single DRG, allows comparison with current performance measurement approaches which ignore quality or outcome variables (mortality rates here). Correction of size-biasing inherent in use of

inverted rates alongside admissions is empirically demonstrated in comparison of efficiency measured under the third specification, with that of the second specification.

Dimensionality problems inherent in specifying quality as an additional output to activity are illustrated in comparison of efficiency estimated with the fourth, relative to the third, specification. Only the fourth specification provides disincentives for low quality (high disutility bearing event rates) as well as incentives for high quality care in technical efficiency measurement, without restrictions for multiplier bounds (assurance regions).

4.6 Results – specifying disutility events as an output

Table 4.6.1 compares input-orientated technical efficiency scores, under constant and variable returns to scale, and economic efficiency scores for each output specification, assuming strong disposability of inputs, with four alternate specifications of output as outlined in section 4.3. To allow demonstration of size-biasing in including quality under specification 2, hospitals are ordered by size (number of admissions for DRG e62a), and the mortality rate for each hospital is included.

Table 4.6.1: Comparing out specifications of mortality rates for DRGe62a

Technical efficiency (Fi), under constant (C) returns to scale and economic efficiency (Oi) with output specifications of:

(1) admissions only
(2) admissions and 1/mortality rate
(3) admission and (1/mortality rate) × admissions
(4) survivors only

	Fi(y,x | C)				Oi(y,x,w | C)				mortality
DMU	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	**rate**
1	0.98	1.00	0.98	0.71	0.74	0.78	0.74	0.53	40%
2	0.67	0.75	0.70	0.69	0.39	0.42	0.39	0.35	25%
3	0.49	1.00	1.00	0.55	0.45	0.83	0.50	0.49	8%
4	0.46	0.90	0.52	0.52	0.29	0.54	0.33	0.32	7%
5	1.00	1.00	1.00	0.87	0.70	0.71	0.70	0.50	40%
6	0.70	1.00	0.99	0.89	0.44	0.81	0.52	0.48	6%
7	1.00	1.00	1.00	0.84	0.87	0.88	0.87	0.67	35%
8	0.79	0.90	0.79	0.81	0.60	0.64	0.61	0.61	14%
9	0.62	0.76	0.67	0.67	0.49	0.52	0.50	0.50	13%
10	0.80	1.00	1.00	0.98	0.54	1.00	0.73	0.61	4%
11	0.67	1.00	0.74	0.72	0.48	0.88	0.64	0.54	4%
12	0.75	0.75	0.75	0.72	0.43	0.43	0.43	0.35	32%
13	1.00	1.00	1.00	1.00	0.59	0.59	0.59	0.43	38%
14	0.49	0.84	0.54	0.52	0.27	0.51	0.39	0.31	4%
15	0.75	0.90	0.86	0.85	0.54	0.58	0.57	0.58	10%
16	1.00	1.00	1.00	1.00	0.58	0.58	0.58	0.52	25%
17	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	9%
18	0.56	0.56	0.56	0.51	0.48	0.49	0.48	0.44	24%
19	0.98	1.00	1.00	1.00	0.79	0.83	0.82	0.83	12%
20	0.91	0.91	0.91	0.82	0.59	0.59	0.59	0.53	24%
21	0.54	0.55	0.55	0.55	0.48	0.49	0.49	0.49	14%
22	0.82	0.82	0.82	0.68	0.74	0.74	0.74	0.65	26%
23	0.78	0.78	0.78	0.73	0.61	0.61	0.61	0.57	21%
24	1.00	1.00	1.00	0.99	0.68	0.68	0.68	0.56	30%
25	1.00	1.00	1.00	1.00	0.79	0.79	0.79	0.74	21%
26	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	17%
27	0.68	0.69	0.78	0.74	0.59	0.61	0.70	0.65	6%
28	0.83	0.83	0.86	0.96	0.46	0.46	0.46	0.45	18%
29	0.99	0.99	1.00	1.00	0.68	0.68	0.71	0.71	11%
30	0.75	0.75	0.75	0.59	0.61	0.61	0.61	0.49	32%
31	1.00	1.00	1.00	1.00	0.65	0.65	0.65	0.64	17%
32	0.76	0.76	0.76	0.67	0.53	0.53	0.53	0.46	27%
33	1.00	1.00	1.00	1.00	0.68	0.73	1.00	0.78	3%
34	0.58	0.58	0.60	0.60	0.51	0.51	0.55	0.55	10%
35	0.59	0.59	0.59	0.54	0.48	0.48	0.48	0.44	24%
36	0.98	0.98	0.98	0.94	0.69	0.69	0.69	0.62	25%
37	0.90	0.90	0.90	0.77	0.62	0.62	0.62	0.51	30%
38	0.62	0.62	0.62	0.57	0.52	0.52	0.52	0.47	23%
39	0.77	0.77	0.77	0.64	0.56	0.56	0.56	0.46	31%
40	0.77	0.77	0.77	0.74	0.61	0.61	0.61	0.57	21%
41	0.81	0.81	0.81	0.70	0.64	0.64	0.64	0.54	29%
42	0.76	0.76	0.76	0.71	0.51	0.51	0.51	0.47	21%
43	0.92	0.92	0.92	0.74	0.67	0.67	0.67	0.53	34%
44	0.55	0.55	0.55	0.49	0.47	0.47	0.47	0.41	27%
45	0.78	0.78	0.78	0.69	0.53	0.53	0.53	0.45	28%

4.6.1 Overcoming size-biasing by normalising inverted rates

As predicted theoretically, including the reciprocal of mortality rates specified as an additional output to admissions (comparing specifications 1 and 2) is size-biased, improving efficiency more for hospitals with lower number of admission at the same disutility event rate. For example, for hospitals 6, 10, 11 and 14, with low mortality rates and number of admissions less than 30, economic and technical efficiency under constant returns to scale improved significantly. However for hospitals 27 and 34, with low mortality rates and admissions greater than 60, efficiency either remained the same or barely changed. Specification (3) overcame this size-biasing problem. Relative to ignoring quality in specifying admissions alone (1), each of the hospitals 6, 10, 11, 14, 27 and 34 increased efficiency under specification (3) as expected, given their relatively low mortality rate.

Empirically, use of inverted rates multiplied by admissions in normalising output ratios removes bias by hospital size inherent in the use of inverted disutility event rates without normalising. However, while quality specified as inverted mortality multiplied by admissions provides appropriate relative performance measurement at the same level of mortality, questions remain over relative performance measurement across different levels of disutility.

4.6.2 Non-linearity in valuing of quality of care with inverted disutility event rates

While normalising inverted disutility event rates on activity can avoid size-biasing, the implicit value of avoiding disutility events is not linear across different levels of disutility rate. Specification of an inverted disutility event rate measures output as though the absolute value of reducing the disutility event rate increases, as the disutility rate falls. For example, an absolute reduction in disutility of 1% with 100 admissions leads to a:

1. 1% or 1 unit increase in quality measured at a disutility rate of 100% (from 100 to 101 with 100 admissions);
2. 2% or 4 unit increase in quality at a disutility rate of 50% (from 200 to 204 with 100 admissions) and a;
3. 100% or 5000 unit increase in quality at a disutility rate of 2% (from 5000 to 10000 with 100 admissions).

In general, as rates of disutility events approach zero, inverted rates asymptote to infinity (undefined at 0), while for high rates of disutility bearing events they tend to 1 (at 1 for 100%). If quality of care is specified with inverted disutility event rates, then implicitly there is not an appropriate trade off between cost and quality of care. Dichotomous incentives would be created in performance measurement for high quality-high cost care, and low quality-low cost care.

In rewarding high quality but not punishing low quality, incentives are created for a divergence into high quality-high cost care and low quality-low cost care. Low quality (quality effectively valued at 0) encourages care where marginal benefit is greater than marginal cost, while rewarding high quality care encourages too high a quality of care, given marginal cost and benefits of increasing quality.

The direction of the non-linearity created by this inverted measure also directly opposes the direction of preferences that prospect theory (Kahnemann and Tversky, 1979) suggests, given loss aversion. Loss aversion is suggested under prospect theory to result in higher valuations of perceived losses than gains (from a position of endowment), and hence willingness to accept loss is valued higher than willingness to pay for equivalent gain. In reviewing empirical studies in health care, Willan, O'Brien and Leyva (2001) found a 2-3 fold greater value of willingness to accept (WTA) health loss than willingness to pay (WTP) for equivalent health gain.

Prospect theory diverges from the predictions of Coase's theorem of the law of one price in the fundamental theorem of exchange. Under the law of one price, ignoring any income effects, willingness to pay (WTP) for gain would equal willingness to accept (WTA) loss, and could be thought of as constant. Mitchell and Carson (1989) posited alternative explanations of the extent of discrepancy between WTP and WTA of:

1) Measurement artefact from contingent valuation where individuals are assigned property rights over public goods.
2) A lack of substitute commodities.

However, Kahnemann, Knetsch and Taylor (1990) demonstrated that the empirical finding of willingness to accept (WTA) loss as being greater than willingness to pay

(WTP) for the same gain persists, even for private goods, and allowing for learning effects in trading.

To reflect preferences for quality of care, if there is any divergence from linear valuation, the marginal value of avoiding disutility event rates should therefore, if anything, be higher rather than lower, as quality of care falls below (or equivalently disutility rates increase above) a threshold level of endowment. Above this disutility event rate, health effects of care across a population can be perceived as a health loss. The implicit diminishing marginal value from reductions in disutility rates with inverted rates can therefore be seen as acting in the opposite direction to preferences in valuing effects, while creating perverse dichotomous incentives for high and low quality of care.

4.6.3 Dimensionality limitations with quality as an additional output

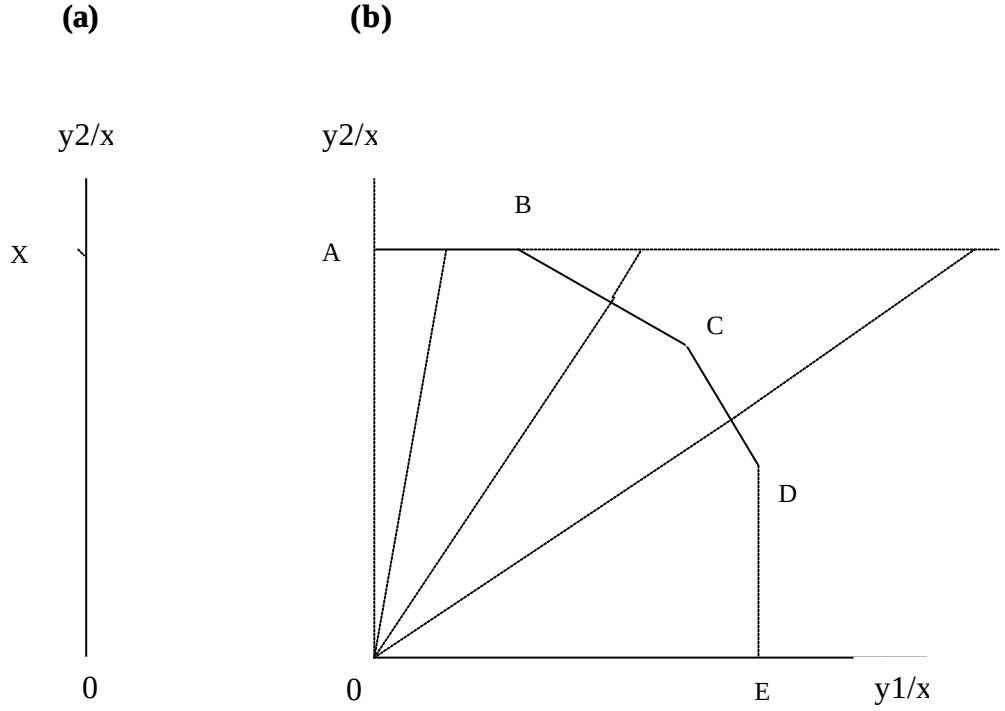
Output specifications of quality, such as inverted disutility event rates, in the absence of relative prices for admissions cannot estimate output-orientated economic, or allocative, efficiency. Therefore, technical efficiency, by default, becomes the performance measure allowing for quality of care, rather than one aspect in decomposition of economic efficiency with an appropriate objective function.

Technical efficiency for any given decision making unit (DMU) is estimated with DEA relative to a frontier of best practice, constructed as a piece-wise linear hull, in radial contraction in an input-orientation, and radial expansion in an output-orientation. Hence, providers (DMUs) are compared with technical efficiency to other providers with similar factor input or output ratios. While this is an advantage of DEA in allowing flexible comparison of providers (DMUs), dimensionality problems of technically inappropriate tradeoffs between inputs or outputs, and quality in particular, arise as a consequence of this flexibility in estimating technical efficiency.

The first set of dimensionality problems arise with DEA, as the addition of a dimension (input or output) to a DEA formulation decreases the ability to discriminate between relative technical efficiency of observed decision making units. In general, for a given set of decision making units, increasing the number of inputs or outputs can only increase average technical efficiency across an industry in comparing performance

(Farrell, 1957). This is a direct result of the dimensionality of input and output space relative to the number of observations (Seiford and Thrall 1990), and a lack of restriction on multipliers in DEA formulations. When an additional output, such as quality, is added to a DEA formulation for a given set of decision making units, both the industry technical efficiency and each decision making unit's technical efficiency can only increase. For example, in figure 4.2, this is illustrated in comparing technical efficiency under constant returns to scale when an output representing quality (y1) is added to efficiency measurement with one input (x) and one output (y2).

Figure 4.2 Dimensionality and technical efficiency adding quality as an output in DEA



The addition of quality as an output variable (y1) can only maintain or reduce the extent of radial expansion (of outputs per unit input under CRS) of any given DMU to a DEA frontier, represented by the single point X in the case of one output and frontier ABCDE for two outputs. For example, for hospitals with low quality of care that have radial expansion projecting onto section AB, technical efficiency will remain the same with addition of y1 (quality). For DMUs projecting onto the frontier BCDE, technical efficiency increases relative to the one output case, given that a radial expansion to the extension of line segment AB is equivalent to that required in the one input, one output case.

An inability to provide disincentives for low quality of care in measuring technical inefficiency with quality specified as an output would not be problematic, if an economic efficiency measure allowing for value of outputs were also estimated. However, neither outputs of admissions or quality estimated with a variable such as inverted mortality rates have prices, and hence revenue or profit efficiency cannot be calculated and performance measurement reverts to technical efficiency.

Dimensionality limitations from adding quality specified as an output (e.g. normalised inverted disutility event rates) separately to quantity (admissions) as an output can be seen in table 4.6.1. Both the third and fourth specifications, in including quality of care, appropriately increase efficiency scores for hospitals with relatively low mortality rates compared with an output specification of admissions alone. However, for hospitals with relatively high mortality rates (for example hospitals 1, 5, 7, 13, 24, 30, 37, 39, 43), only the fourth specification of survival alone allows reduction in technical or cost efficiency relative to admissions alone. Technical efficiency for the third output specification of admissions and normalised inverted disutility event rates face a dimensionality limitation in providing disincentives for low quality in trading off quality and quantity. The fourth output specification of admissions without disutility events (where quality and quantity are not separated) does not face this dimensionality limitation.

4.6.4 Dimensionality, technical efficiency and quality-quantity trade-offs

While a sufficient number of observations relative to dimensions are necessary to enable discrimination, where quality and quantity are separate, this does not ensure appropriate trade-offs occur between quality and quantity of care in estimating technical efficiency with DEA. In including quality as an additional output to activity in performance measurement, unacceptable ratios and trade-offs can be present in measuring technical efficiency in the absence of restrictions on weights. The reporting of slacks, as potential reductions in inputs or increases in outputs, holding other factors constant at the target on the frontier (such as projections onto AB or DE in figure 4.2) in combination with technical efficiency scores, can partly allow for this. However, reporting of slacks does not prevent extreme ratios on the frontier. Slacks are only measured relative to regions of frontier beyond observed best practice, which can include 'corner solutions'. A DMU can be included on the frontier with extreme factor ratios that do not reflect technically acceptable ratios between inputs and/or outputs. These corner solutions are allowed as a consequence of flexibility in output multipliers, where decision making units can specialise in production of an output (e.g. admissions or quality) such that the number of comparators they face reduces, in the extreme, to just themselves.

A second set of dimensionality problems therefore relate to incentives for corner solutions, where performance measures with multiple output providers (DMUs) can limit comparators and be measured as more technically efficient, if they specialise in one output at the expense of all others. Incentives for corner solutions to specialise in one output are present in multiple output specifications with either technical efficiency or cost efficiency measures with input-orientated DEA, as neither imposes restrictions on weight for output factor ratios.

A third set of dimensionality problems with multiple outputs specifications in hospitals relates to output slacks. If price data for inputs is available, measuring cost efficiency allows input slacks and technically inappropriate ratios to be included as allocative inefficiency, as suggested by Ferrier and Lovell (1990:235), and outlined in appendix 4.1. However, with multiple outputs, output slacks can arise in addition to input slacks. For example, hospitals with the lowest and highest output ratios must either be on the DEA frontier or have output slacks, given any combination of other hospitals cannot have as extreme an output ratio. Output slacks should therefore be reported with technical or economic efficiency if a multiple output specification were used, even if input-orientated.

In public hospitals, the absence of prices for outputs (and hence revenue efficiency) implies an inability to subsume output slacks into an allocative efficiency measure. This is clearly a dimensionality limitation of a multiple output DEA specification. In modelling quality as an output, complete freedom of choice of weights implicit in DEA formulations is, as Thanassoulis, Boussofiane and Dyson (1995:594) state: "...not wholly appropriate where output quality measures are present in the model".

An inability to create disincentives for lower quality, incentives for corner solutions and the need to report slacks can be seen as dimensionality limitations of DEA specified with quality as an additional output, due to the absence of weight restrictions on output factor ratios.

4.6.5 Assurance regions

The dimensionality problem of including quality as an extra variable may be mitigated against to some extent, with modification of the DEA model to restrict weight flexibility, to create what Thomson et al. (1986, 1990, and 1992) describe as 'assurance regions'.

Restrictions on multiplier bounds mitigate against corner solutions by restricting the shape of the frontier to regions reflecting restricted input and output trade-offs. Assurance regions can therefore both remove firms from the DEA frontier in regions with multipliers outside acceptable ranges and increase the relative inefficiency of firms who had these frontier firms as peers. In an input-orientation, restrictions on output weights affect both technical and cost efficiency.

In the case of public hospitals, imposing multiplier bounds between output variables of quality, measured as inverted disutility rates multiplied by admission, relative to admission, is equivalent to placing restrictions on disutility bearing outcome rates. The restrictions in use of multiplier bounds could therefore be used to create disincentives for high rates of disutility bearing events in calculating both technical and economic efficiency. Hospital technical efficiency would then be able to decrease with the addition of quality as an output variable.

The justification for use of assurance regions is usually based on bounds for technically feasible input ratios in production. However, in using technical efficiency for performance measurement, with quality represented as an output alongside admissions, assurance regions can be justified on grounds of appropriate standards of care. In the case of quality of care, represented by disutility event rates, the choice of restrictions on disutility bearing rates could be based on:

(1) a statistical procedure such as 2 standard deviations above the mean rate of disutility events or;

(2) comparison with next best alternative treatment.

The second option arises from prospect theory (Kahnemann and Tversky, 1979). As outlined in section 4.6.2, loss aversion relative to a position of endowment results in higher valuations of perceived losses than gains. Assurance regions in DEA could

therefore be determined, by limiting the rate of disutility bearing events, to below that at which the net effect of hospital care could be seen as a loss relative to next best alternative care.

Thanassoulis, Boussofiane and Dyson (1995) suggested the need for, and demonstrated use of, assurance regions and related methods in including positive quality indicators as additional outputs, in a DEA study of perinatal care. Specification of assurance regions could potentially allow for some dimensionality limitation in performance measurement. However, choice of assurance regions is somewhat arbitrary and will only influence incentives up to whatever threshold level of trade-off is chosen. Specification of quality as an output, integrated with activity alone, may be considered a better option, both theoretically and empirically, in overcoming dimensionality and linearity problems and in satisfying parsimony.

4.6.6 Specifying admissions without disutility events

In our case example, specifying a single output as survivors (admissions without deaths), there is an explicit trade-off between quality and quantity for all hospitals, whether they have low or high mortality rates. For hospitals with relatively high mortality rates (for example hospitals 1, 5, 7, 13, 24, 30, 37, 39, 43 in table 4.6.1), a specification with survival alone appropriately reduced technical and economic efficiency scores relative to specifications without quality included.

Relative to a specification of admissions alone as an output, survival alone as an output both rewards (provides incentives for) low mortality rates and punishes (provides disincentives for) high mortality rates. In general, admissions free of disutility events as an output specification avoids dimensionality problems of separate specification of activity and quality and ensures a trade-off between disutility events and admissions. To the extent it ensures a trade-off, it may therefore be considered more appropriate, in incorporating quality in estimating technical efficiency for performance measurement (given the inability to measure economic efficiency), than other specifications with quality as an additional output.

However, specifying admission free of disutility bearing events implicitly assumes that admissions with disutility bearing events are valued at 0. Whether this is a reasonable

assumption will depend on the type of disutility bearing event. Assuming the value of avoiding a death is considered greater than that of an admission, performance measurement and incentives are at least moved in the appropriate direction. No perverse incentives are created for quality of care in comparison with ignoring quality (using admissions alone as outputs), and the value of quality in avoiding mortality starts to be recognised.

In general, a specification of admissions without disutility events is only likely to provide an appropriate value with the single disutility events of readmission for the same DRG. For serious health effects such as mortality, admissions with these events are net disutility bearing, and thus a specification such as survivors alone, while moving in the right direction from admissions alone, understates the disutility of these events. Conversely, for minor iatrogenic events a 0 value may overstate the value of disutility relative to that of admissions. Under this specification the question is also raised of how to allow for multiple disutility events, which, depending on clinical activity might include mortality, different forms of morbidity and readmission.

For a single health related disutility event an output of admissions with the absence of disutility events (for example survivors as the absence of mortality) represents average cost effectiveness (cost per survivor). Even if health related disutility were measurable, output under this specification would become average health related utility at separation, and hence the underlying objective function would represent average cost effectiveness. In measuring relative hospital performance, average cost effectiveness does not, however, appropriately allow for the incremental effect of health care on health (given the stock nature of health) and the non-traded nature of health effects specific to patients treated.

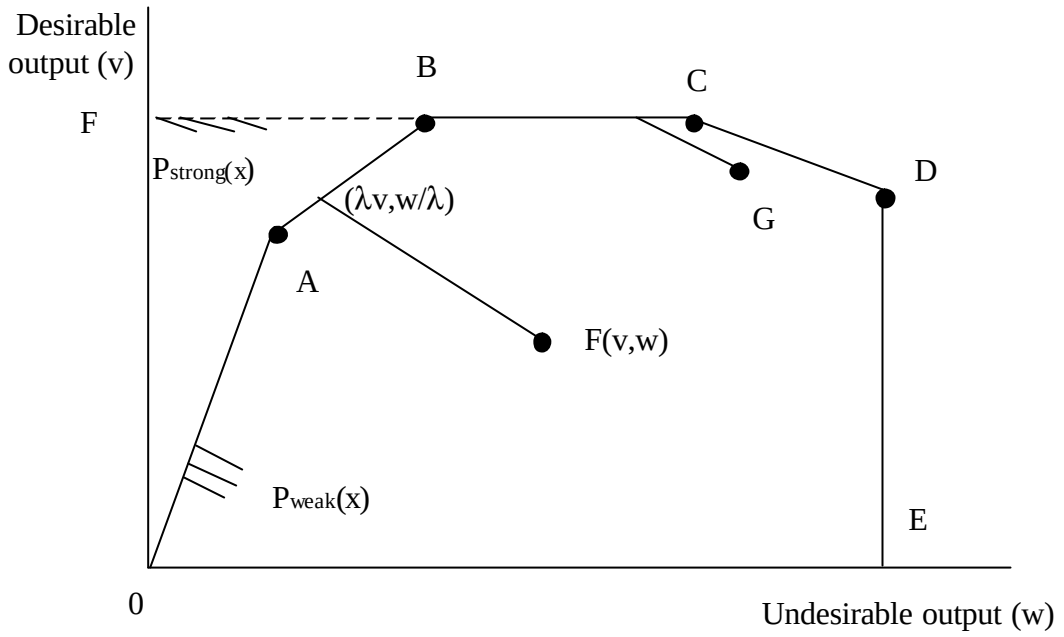
The incremental nature of health effects of care implies that a lower cost-effectiveness ratio, attributable to lower costs but also worse outcomes, will not necessarily be preferred to a higher ratio. The non-traded nature of health effects implies a lower cost per incremental effect will not necessarily be preferred, given the inability to factor up health gains in given populations. In combination, the incremental and non-traded nature of health effects of care imply that relative performance of providers, as for relative performance of technologies in health technology assessment, should consider

incremental effects relative to the incremental costs. The distinction between relative performance based on average cost effectiveness and that of incremental cost effectiveness relative to a threshold, implicit in maximising net benefit, are considered in detail in section 5.7.2.2.

4.7 Hyperbolic specification of disutility events as undesirable outputs

An alternative approach to specifying disutility event rates as an output is the method of Färe, Grosskopf, Lovell and Parsuka (1989), modelling undesirable output as a weakly-disposable negative output. Using this method, technical efficiency is estimated in equi-proportionally increasing desirable outputs, and reducing undesirable outputs, relative to a frontier constructed with undesirable outputs as weakly disposable, and desirable outputs as strongly disposable. The equi-proportional expansion and contraction to a frontier is achieved using a hyperbolic form of DEA (Färe, Grosskopf and Lovell, 1985). Figure 4.3 illustrates a hyperbolic frontier under weak and strong disposability of undesirable outputs as in Färe, Grosskopf, Lovell and Parsuka (1989).

Figure 4.3 Hyperbolic DEA with equi-proportional expansion of desirable outputs and contraction of undesirable outputs



This hyperbolic method has been used in allowing for environmental pollution as an undesirable by-product in production under government regulation of marketed outputs such as electricity. However, adapting this method to hospital performance measurement, with quality of care represented by rates of disutility events, is problematic. Problems are faced in:

(1) separating desirable and undesirable outputs where disutility event rates represent quality of care and admissions throughput;

(2) interpreting technical efficiency as a performance measure where the hyperbolic frontier has sections with increasing undesirable outputs and reducing desirable outputs; and

(3) estimating a shadow price for avoiding disutility events where admissions *per se* are not sold in a market.

4.7.1 Problems of separating 'desirable' and 'non-desirable' outputs with quality differentiated care

The initial problem in adapting the hyperbolic method is separating desirable and undesirable outputs with quality differentiated care represented by disutility event rates. Admissions and disutility bearing events are not theoretically separable as marketed

desirable outputs and undesirable outputs in the same way that, for example, electricity generation and pollution are in the environmental studies where the approach has been used.

Admissions *per se* do not necessarily have intrinsic value in hospital care, given the derived nature of demand for health care, as outlined in section 1.3. It is the relative effect of care in reducing disutility bearing effects such as mortality, morbidity, readmission, functional limitation or disutility that can be considered as valued (although see sections 5.7.2.7 and 8.2.1 for a discussion of including the value of process aspects of care). Quality represented by disutility event rates is an intrinsic characteristic of admissions, rather than a by-product separable from admissions as a 'desirable' output.

4.7.2 Problems of interpreting technical efficiency as a performance measure in the absence of economic efficiency

A second set of problems with the hyperbolic method relate to interpreting technical efficiency relative to regions of the frontier where disutility event rates and cost per admission are increasing. Unless there are external constraints on rates of disutility events, technical efficiency relative to a production possibility frontier (PPF) with a desirable and undesirable output is only meaningful as a performance measure while the frontier is non-negatively sloped.

In assessing relative performance, if both desirable and undesirable outputs were modelled as strongly disposable, then no region of the frontier EDCBF in figure 4.3 would be positively sloped. Hence, technical efficiency relative to this frontier would not allow meaningful representation of performance in maximising desirable, and minimising undesirable, outputs. Under weak disposability of undesirable outputs, the frontier OABCDE proposed by Färe, Grosskopf, Lovell and Parsuka (1989), is appropriately positively sloped in region OAB. However, in the negatively sloped section CDE, technical efficiency is still not meaningful as a performance measurement, assuming costs and effects of care are determined by hospitals (endogenous). For example, in figure 4.3, a producer at D is on the frontier, as there is no further equi-proportional increase in desirable outputs and decrease in undesirable outputs possible, but clearly cannot represent best practice. For the same inputs, D can

produce more desirable outputs with less undesirable outputs, for example at G off the frontier. However technical efficiency under the hyperbolic approach is measured to be greater at D than G.

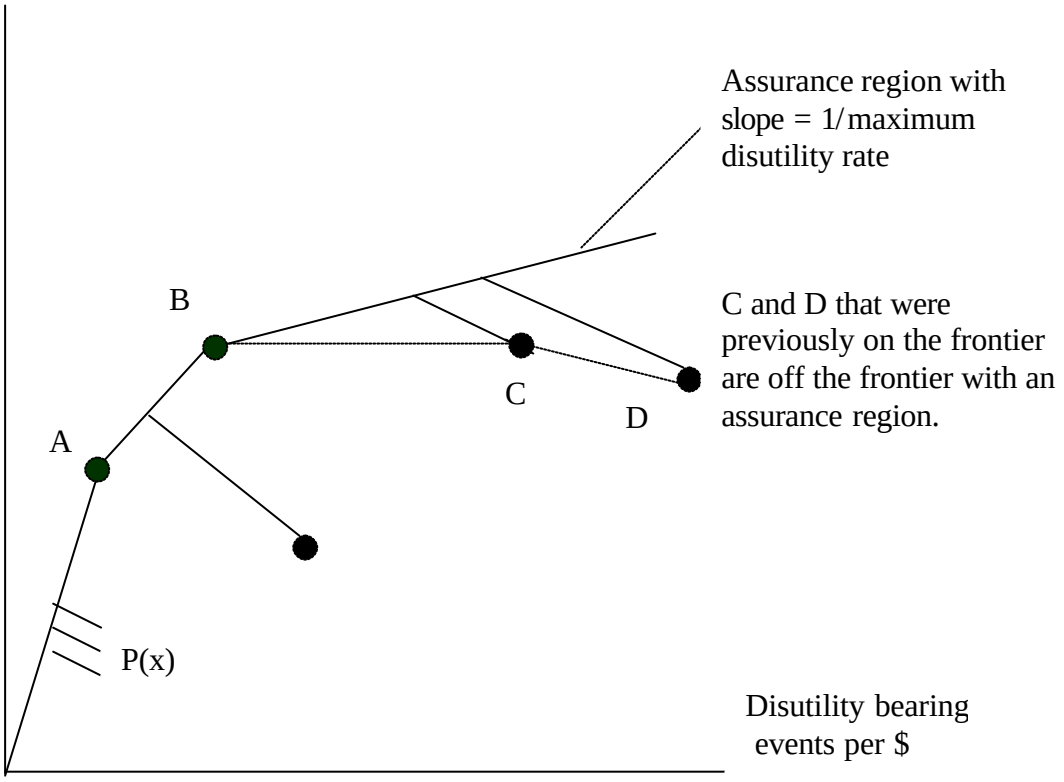
A wedge between technical efficiency and performance measurement in modelling disutility events as undesirable outputs might not be considered problematic if economic efficiency, reflecting the value of outputs, were also able to be calculated. Perversity in provider's behaviour in producing on regions of the frontier increasing disutility events and reducing utility bearing could then be subsumed into allocative efficiency as the residual of economic and technical efficiency. This argument follows similar logic to that of Ferrier and Lovell (1990:235), that slacks representing inappropriate output mixes should be included in allocative (rather than technical) inefficiency.

However, a lack of prices for 'desirable' outputs in hospitals prevents estimating economic efficiency from an output-orientation. Hence, perverse implicit values inherent in technical efficiency measured relative to regions on the frontier such as CDE in Figure 4.3 become a problem for performance measurement. Without any restriction on intensity weights, regions equivalent to an incremental disutility event rate of 100% (comparing B and C in figure 4.3.), or even greater than 100% mortality (comparing say D and C) are permitted.

Only if there was a constraint on firms such that disutility event rates were assumed to be determined outside of the provider's control, rather than representing quality of care, could technical efficiency relative to the region CD have a meaningful interpretation for performance measurement. However, if disutility event rates are adjusted for patient risk factors (environment)[1], then, ignoring the role of chance, they can be assumed to be under the control of the service provider and represent quality of care.

[1] A clinical activity (DRG) level of efficiency analysis allows adjustment of disutility event rates for patient risk factors within DRG, as discussed in chapter 3 and considered in detail in section 5.7.

Figure 4.4 Assurance regions for hyperbolic specification of quality



To allow a non-negative sloped PPF and meaningful relative performance measurement representing disutility events as indicators of quality of care[2], assurance regions at a maximum acceptable disutility bearing event rate, as depicted in figure 4.4, may be seen as necessary. Thanassoulis, Boussofiane and Dyson (1995) suggested the requirement for such assurance regions when using DEA to estimate technical efficiency, including quality with utility-bearing variables.

However, assurance regions such as that depicted in figure 4.4 cannot be constructed using DEA under the hyperbolic specification, with equi-proportional expansion of strongly disposable desirable outputs and contraction of weakly disposable undesirable outputs. The region BCD forms a central part of the hyperbolic production possibility

[2] Assuming differences in patient risks have been adjusted for, such that rates of disutility events represent quality of care

frontier, which cannot be influenced by disposability assumptions or weight restrictions. The hyperbolic method therefore faces intractable problems in identifying technical efficiency with a meaningful interpretation as a performance measure, unless all providers lie in the region OABC.

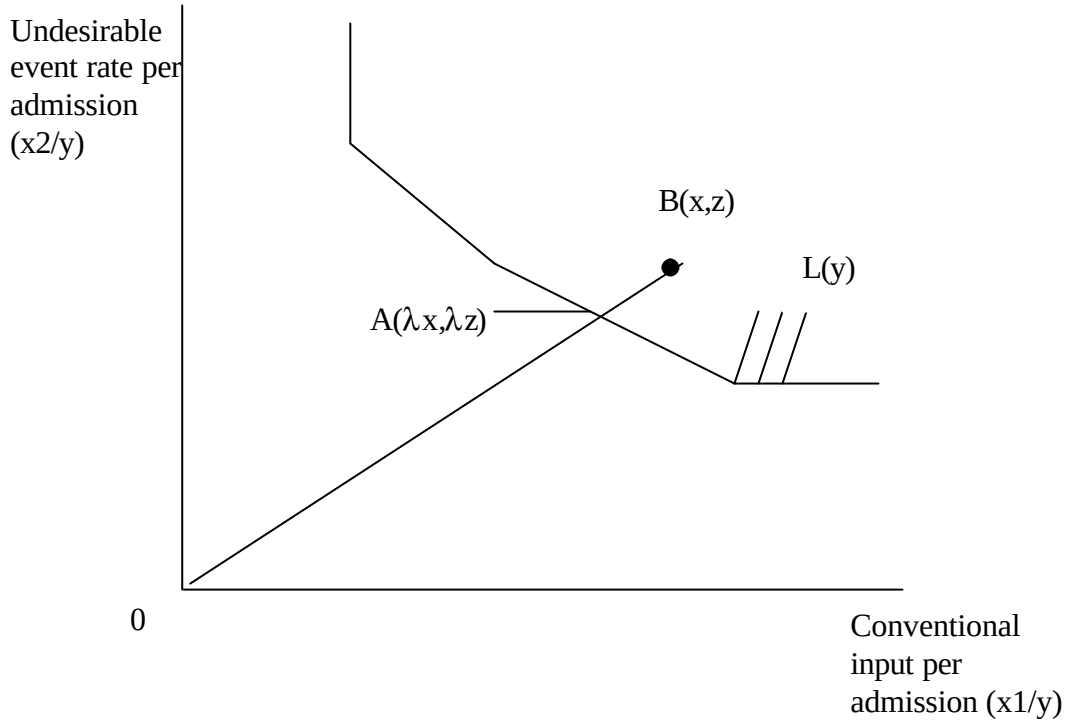
4.7.3 Estimating shadow prices in the absence of prices for admissions

The third problem faced in translating the hyperbolic method to performance measurement in hospitals relates to the absence of a market price for admissions *per se*. The method of Färe, Grosskopf, Lovell and Yaisawarang (1993) allows the shadow or marginal abatement price of pollution, under the hyperbolic specification, to be estimated, using market prices for desirable outputs. However, no market prices exist for admissions *per se* measured as a 'desirable output' in applying the hyperbolic method to hospitals.

A shadow price for disutility events as 'undesirable outputs', can only be estimated in monetary terms relative to cost of admissions. A hyperbolic is therefore suggested, where admissions are held constant specified as a desirable output, and costs and disutility events are specified as inputs and equi-proportionally reduced. This naturally occurs when disutility events are specified as strongly disposable inputs. . Specified as an input, disutility events can be directly and intuitively traded off relative to input use or costs, and shadow price is identifiable. An input specification also allows problems of separability to be overcome. In considering value of health-related quality of care, activity (admissions) can be seen as a scaling factor, rather than as an output separate to effects of care

With undesirable events specified as a strongly disposable conventional input, convexity is ensured and hence, *ceteris paribus*, a greater rate of undesirable events cannot increase technical efficiency (?=OA/OB), as illustrated in figure 4.5

Figure 4.5 specifying undesirable events as an input



4.7.4 Congestion efficiency with input versus hyperbolic output specification of disutility events

If disutility events (e.g. mortality) are modelled as an input, congestion efficiency can also be estimated as the residual of technical efficiency under weak and strong disposability of these inputs. Congestion efficiency with an input specification of disutility events represent increasing resource requirements (costs) of care when disutility event rates increase above (quality of care falling below) a certain level, and can result in expected increased resource requirements and costs of care (from treating disutility events within admission).

In comparison, using the method of Färe, Grosskopf Lovell and Parsuka (1989), with undesirable events specified as outputs, the residual of technical efficiency, specified with weakly and strongly disposable undesirable outputs, represents increase in marketed outputs possible if constraints of reduction in disutility events were removed (pollution abatement). However, congestion efficiency under this hyperbolic specification is not meaningful in hospitals with admissions specified as disutility events and disutility events representing quality of care. Congestion efficiency under this hyperbolic output specification represents the proportional increase in admissions possible if there were no constraints on disutility event rates, failing to recognise the derived nature of demand for health services.

In hospitals, congestion efficiency related to increased expected resource use when disutility rates increase from an input specification is more meaningful than congestion of admissions without constraints on quality of care. The comparison between input and output specifications is explored further in chapter 7, section 7.2.

In identifying technical efficiency which can be interpreted as a performance measure, an input specification therefore has clear methodological and empirical advantages over an output specification as a weakly disposable negative output. As chapter 5 demonstrates, an input specification of technical efficiency also, however, allows an economic efficiency measure corresponding to an underlying appropriate objective function of net benefit maximisation.

4.8 Summary

In this chapter, methods to measure performance incorporating quality of care through disutility events, specified as outputs in estimating technical, cost and allocative efficiency have been critically assessed. The relative characteristics of these output specifications, and particularly their inability to allow appropriate performance measurement incorporating quality of care, have been illustrated with a case example using Australian data for admissions, cost and disutility event rates (mortality) at a DRG level. While admission and inverted disutility event rates were demonstrated theoretically and empirically as size-biased, two specifications have been identified which, in including quality, do not bias efficiency by size of hospital, namely:

(1) admissions free of disutility bearing events and;
(2) admissions and inverted disutility event rate multiplied by admissions.

Of these methods, only admissions free of disutility bearing events provides disincentives for low quality of care as well as incentives for high quality of care, and implicitly allows a linear valuation of disutility bearing events. However, this linear valuation of avoiding disutility events is not flexible, being fixed at a 0 value for admissions with disutility events. This specification also does not allow for the stock or non-traded nature of health in relative performance measurement, at best representing an underlying economic objective of average cost effectiveness.

The alternative hyperbolic approach of Färe, Grosskopf Lovell and Parsuka (1989), equi-proportionally increasing desirable and reducing weakly disposable undesirable outputs, while appropriately used in other settings, has been demonstrated as problematic in representing disutility events as quality of care. Activity and quality are not separable without making inappropriate assumptions such as admissions *per se* are desirable outputs. Technical efficiency, measured relative to regions on the frontier where disutility events increase and 'desirable outputs' fall, is inappropriately interpreted as performance where disutility events reflect quality of care. This is particularly problematic, as technical efficiency becomes the de-facto measure of performance allowing for quality of care under the hyperbolic approach, given an inability to measure output-orientated economic efficiency. Finally, shadow prices of disutility events have been shown to not be estimable using the method of Färe,

Grosskopf, Lovell and Yaisawarang (1993), in the absence of prices for admissions as a separable 'desirable output'.

In general output-orientated specifications of disutility events representing quality of care do not allow a flexible or appropriate trade-off between cost and value of quality of care in relative performance measurement. They do not account for the floor and non-traded nature of health in determining trade-offs, as in net benefit maximisation in health technology assessment. The analysis points to advantages of an approach specifying disutility events as inputs, in allowing meaningful estimation of technical and congestion efficiency and shadow prices for avoiding disutility events relative to costs of inputs. Chapter 5 develops an input specification of disutility events, which allows flexible valuation of disutility events in ordering performance consistent with maximising net benefit per admission.

Chapter 5: The correspondence theorem - specifying quality with disutility events as inputs

5.1 Overview

The current hospital performance measure of cost per case-mix adjusted separation creates perverse economic incentives for reduction in quality of care when used in benchmarking, peer identification or funding. Policy and decision makers recognize the need to allow for quality or care to overcome these perverse incentives, but require a method with appropriate tradeoffs between costs and value of quality of care.

In health technology assessment an objective of maximising net benefit per person as the value of incremental health effects less incremental costs of technologies allows decision makers to trade-off the cost, and value, of quality of care. Analogously, in hospitals an objective of maximising net benefit per admission would allow a trade-off between the cost, and value, of quality of care.

In efficiency measurement, requirements for non-negative inputs and outputs to enable ratio measurement with radial contraction are not satisfied under a direct net benefit formulation. However, a linear translation of net benefit is demonstrated to allow ordering of performance consistent with maximising net benefit, using non-negative inputs of costs and disutility events (such as mortality, morbidity, readmission or functional limitation) and outputs of admissions. A one-to-one correspondence is demonstrated between maximising net benefit and minimising cost plus disutility events, valued at the decision maker's willingness to pay (WTP) thresholds as in net benefit, where:

1. disutility event rates capture the effects of care and;
2. hospitals face the same comparator (differences in patient populations expected cost per admission and disutility event rates are adjusted for).

Application of the correspondence to measuring relative performance, peer identification and benchmarking is illustrated using data envelopment analysis with routinely collected

data at a clinical activity (DRG) level. The method is shown to allow conditioning of relative performance on the decision maker's value of avoiding disutility events, identification of regions of value of effects of care over which technically efficient hospitals are peers, and the shadow price for quality in current industry behaviour. Comparisons with current relative performance measurement and peer identification, based on cost per admission, are demonstrated by including a zero value for quality. Where the WTP to avoid disutility events is known, a decomposition of economic efficiency into technical efficiency minimising costs and disutility per admission, and residual allocative efficiency, allows a story to be told of sources of inefficiency, consistent with net benefit maximisation under correspondence conditions.

Satisfying correspondence conditions allows performance measurement corresponding with net benefit maximisation, but also provides a framework for explicitly addressing incentives for cream-skimming and cost- and event-shifting. The linear nature of the correspondence allows simple inclusion of multiple effects, including those beyond separation in satisfying the condition of coverage of the effects of care with disutility events and in preventing economic incentives for cost-shifting and event,-shifting. At a DRG level, decision analytic methods can be employed to comprehensively identify effects within and beyond separation. Perceived utility bearing aspects of care can also be reframed as disutility event rates, where they can be measured as not meeting standards of care, functional limitation or disutility directly.

To allow the common comparator assumptions to be met and incentives for cream-skimming avoided, methods of adjusting for differences in expected disutility event rates and costs at a DRG level, given patient risk factors, are identified. Approaches examined include standardisation of cost and effects prior to performance measurement, use of expected costs and disutility events as non-discretionary outputs in DEA and direct regression on a per admission net loss statistic, conditional on willingness to pay. Methods of adjustment for environmental factors, such as teaching status, available technology or function using peer grouping or second stage regression, and capacity, using variable returns to scale formulations, are also examined.

5.2 Maximising net benefit – trading off cost and value of quality

Health technology assessment (HTA) in the UK (NICE, 2001), Australia (ACDHA, 2002) and provinces in Canada (MHO, 1994) use a systematic evidence-based framework that considers the effects, as well as costs, of health care alternatives, in treating defined patient populations. A trade-off between the value of incremental effects and incremental costs is implicitly made in deciding whether an estimated incremental cost-effectiveness ratio is acceptable, or explicitly in determining net benefit. By contrast, in measuring relative economic performance between health care providers such as hospitals, a systematic framework for integrating value as well as costs has not been identified in measuring relative economic performance. This is despite the need, as Smith (2002:146) identifies, for a: "coherent conceptual framework to inform the design of performance measurement systems", and the collection of quality of care indicators in Australia (NHPC, 2000), Canada (Wolfson and Alvarez, 2000) and the UK (NHS, 2002).

5.2.1 Appropriate quality of care: maximising net benefit per admission

Ideally, in measuring hospital performance, we would like to provide economic incentives to encourage quality of care while the marginal value from increasing clinical quality (reducing mortality and/or morbidity) outweighs the marginal cost of increasing clinical quality of care. Analogous to maximising profit per unit output (revenue less costs), given prices of inputs and outputs, we could envisage maximising net benefit (Stinnett and Mullahy, 1998) per admission. Net benefit per admission for a hospital relative to a comparator can be characterised as the monetary value of incremental health effects less incremental cost per admission:

$$NB_i = k(E_i - E_{comp}) - (C_i - C_{comp}) \qquad (5.1)$$

Where:

NB_i is the net benefit per admission for hospital i $(i = 1..n)$ relative to a comparator $(comp)$;

k represents a decision making threshold value for incremental effects, characterised as willingness to pay (WTP);

E_i and C_i are the effect and cost per admission for hospital i and;

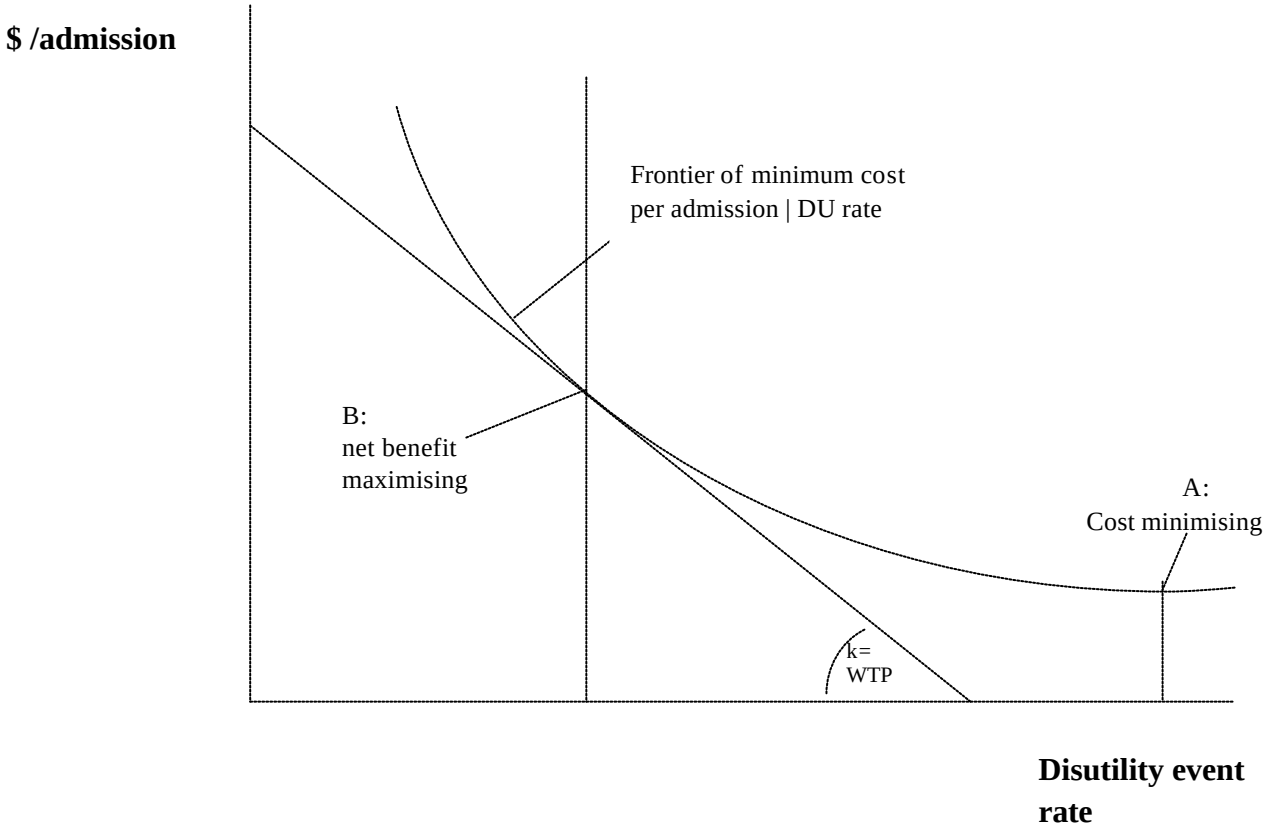
E_{comp} and C_{comp} are the effect and cost per admission for a (unknown) comparator.

5.2.2 Advantages of maximising net benefit versus minimising cost per admission

Measuring performance with cost per separation (whether at a DRG level or case-mix adjusted at an aggregate level), economic incentives are implicitly created for hospitals to reduce quality of care to a level minimising costs per admission. In ignoring effects of quality of care, quality of care only begins to be valued positively if incremental expected cost from reducing quality becomes positive. Economic incentives to reduce quality in identifying peers and benchmarks based on minimising cost per admission, are present while intra-marginal costs across patients of lower quality of care are greater than expected infra-marginal cost increases (from treating higher resulting rates of disutility events within admission). Unless infra-marginal cost savings of increased quality continually outweigh intra-marginal costs of increasing quality, economic incentives are created for quality below that of evidence based medicine (maximising net benefit per admission).

The distinction between cost minimising and net benefit maximising quality of care is shown graphically in figure 5.1, where effects of care are represented by disutility event rates (such as mortality rate), and valued at the decision maker's value of willingness to pay to avoid disutility events (k).

Figure 5.1 Cost minimisation versus net benefit maximisation



A priori, as quality of care increases (expected disutility event rate reduces) diminishing returns to resources and costs of care are eventually expected. The expected marginal cost of reducing disutility event rates increases as the rate of disutility events approaches 0. Characteristically then, in figure 5.1, while cost minimising occurs at point A, for a value of avoiding disutility events of k, net benefit maximising quality occurs at point B. Only if expected infra-marginal cost increases, from treating higher rates of disutility events within-admission, were continually greater than intra-marginal costs across patients of lower quality of care, would point A and B be the same.

While industry behaviour may not reach a level of quality corresponding to minimum cost per admission at A, this is not due to lack of economic incentives provided by performance measured with cost per admission or funding based on case-mix payments per admission. It is the presence of health objectives in provider behaviour, such of those from the Hippocratic oath, professional standing in medical practice and constraints on behaviour from hospital accreditation or access to technology informed by health technology assessment, which in practice may mitigate against a minimum cost per admission level of quality. This degree of opposition will, however, vary in any given activity in any hospital, to an extent depending on local circumstances and, as Harris (1977) suggests, in particular, on the relationship between hospital administrators with an objective of minimising cost per admission under case-mix funding arrangements and performance measures and medical staff, with an objective of health maximisation.[1]

However, if performance measurement and funding for hospitals could be undertaken consistent with maximising net benefit per admission, service providers could be held economically accountable for their cost and effects of care. Incentives could be created that actively support evidence-based medicine, reflecting an appropriate trade-off between the cost and value of effects of quality.

[1] Harris's model of internal hospital organisation is considered in detail in chapter 7, section 7.3 in discussion of policy implications of the correspondence theorem in relation to internal organisation of the hospital, as well as the hospital as an agent.

5.2.3. Finding a specification consistent with net benefit maximisation

In finding a specification for performance consistent with net benefit, the incremental nature of net benefit calculated relative to a comparator needs to be allowed for. Additionally, if efficiency measurement methods based on ratio measurement are to be employed, inputs and outputs need to be non-negative variables. In comparing relative performance between hospitals, the incremental nature of net benefit can be overcome, where hospitals face the same (even if unknown) comparator[2]. However, direct use of the net benefit formulation of equation (5.1) in relative performance measurement violates the requirement that inputs or outputs cannot take negative values. A transformation of net benefit to allow positive measurement of inputs and outputs and consistency with net benefit ordering may, however, be possible.

5.3 The Net Benefit Correspondence Theorem

A one-to-one correspondence exists between ordering hospital performance maximising net benefit per admission and minimising the sum of cost and rates of disutility bearing events per admission, valued in monetary terms as in net benefit, under assumptions that:

1. *effects of care can be represented by disutility event rates, and;*
2. *hospitals face the same comparator, (differences between expected costs and disutility event rates in patient populations treated are adjusted for).*

5.3.1 Proof - Case 1: single effect and a common comparator

Consider a bilateral comparison between hospitals (i, j) where health effects of care can be attributed to the rate of admissions without a single type of disutility event, for example survival rate (E). Let the associated decision maker's value for willingness to pay to avoid disutility events (mortality) be k, and cost per admission, C. Then from (5.1) the net benefit per admission (NB) for hospital i $(i = 1,..,n)$ relative to a comparator $(comp)$ can be expressed as:

$$NB_i = k \times (E_i - E_{comp}) - (C_i - C_{comp}) = k \times E_i - C_i - (k \times E_{comp} - C_{comp}). \qquad (5.2)$$

[2] This condition is not as restrictive as may first appear. If expected cost and effects (e.g. mortality rate) differ across patient populations treated in hospitals, where these differences can be adjusted for, relative comparison can be undertaken as if hospitals faced the same comparator.

Without loss of generalization (ordering is arbitrary in establishing a correspondence), let

$$NB_i > NB_j .$$

Then all comparator cost and effect terms in (5.2) cancel, in comparing two hospitals with the same comparator under the second assumption.

$$\Leftrightarrow k \times E_i - C_i > k \times E_j - C_j . \quad (5.3)$$

If we multiply both sides of (5.3) by minus 1, the sign changes and we translate from maximising net benefit per admission to minimising net loss per admission.

$$\Leftrightarrow C_i - k \times E_i < C_j - k \times E_j . \quad (5.4)$$

To allow performance measurement with positive arguments for all inputs and outputs, we add k to both sides of the equation and re-arrange with common factors.

$$\Leftrightarrow C_i + (1 - E_i) \times k < C_j + (1 - E_j) \times k . \quad (5.5)$$

Now, if E is rate of admissions without disutility events (survival rate), then $1 - E$ is the rate of admissions with disutility events (mortality rate) or DU.

$$\Leftrightarrow C_i + DU_i \times k < C_j + DU_j \times k . \quad (5.6)$$

QED (single effect, common comparator case)

5.3.2 Proof - Case 2: single effect and differences in expected costs and effects

Consider a bilateral comparison between hospitals (i, j) where incremental effect per admission relative to comparators for each hospital ($comp_i, comp_j$) can be attributed to the difference in a single effect, say survival rate (E). Let the associated decision maker's value for willingness to pay to avoid disutility events (deaths) be k, and incremental cost per admission be C. Then from (5.1) the net benefit per admission (NB) for hospital i ($i = 1,..,n$) relative to its comparator can be expressed as:

$$NB_i = k \times (E_i - E_{comp_i}) - (C_i - C_{comp_j}) = k \times E_i - C_i - (k \times E_{comp_j} - C_{comp_i}) . \quad (5.7)$$

Without loss of generalization, let $NB_i > NB_j$.

$$\Leftrightarrow k \times E_i - C_i > k \times E_j - C_j + [(C_{comp_j} - C_{comp_i}) - (E_{comp_j} - E_{comp_i})] \quad (5.8)$$

If we multiply both sides of (5.8) by minus 1, the sign changes and we translate from maximising net benefit per admission to minimising net loss per admission.

$$\Leftrightarrow C_i - k \times E_i < C_j - k \times E_j - [(C_{comp_j} - C_{comp_i}) - (E_{comp_j} - E_{comp_i})] \quad (5.9)$$

To allow performance measurement with positive arguments for all inputs and outputs, we add k to both sides of the equation and re-arrange with common factors.

$$\Leftrightarrow C_i + (1 - E_i) \times k < C_j + (1 - E_j) \times k - [(C_{comp_j} - C_{comp_i}) - (E_{comp_j} - E_{comp_i})]. \quad (5.10)$$

If E is the rate of admissions without disutility events (survival rate), then $1 - E$ is the rate of admissions with disutility events (mortality rate) or DU.

$$\Leftrightarrow C_i + DU_i \times k < C_j + DU_j \times k - [(C_{compj} - C_{compi}) + (DU_{compj} - DU_{compi})]. \quad (5.11)$$

In relative comparison between hospitals with different populations, differences in comparator costs and disutility event rates can be represented by expected differences in costs and disutility event rates.

$$\Leftrightarrow C_i + DU_i \times k < C_j + DU_j \times k - (E[C_j - C_i] + E[DU_j - DU_i]), \quad (5.12)$$

where:

$E[C_j - C_i]$ is the expectation of the cost difference per admission between hospitals i and j, and;

$E[DU_j - DU_i]$ denotes expected difference in disutility event rate between hospitals i and j, given differences in patient risk factors.

Therefore under the second assumption, if differences in expected costs and disutility events are adjusted for, there is a one-to-one correspondence with maximising net benefit.

QED (single effect, adjusting for difference in baseline risk of costs and effects)

5.3.3 Proof - Case 3: multiple disutility bearing events and a common comparator

Let all potential combinations of disutility events observed in patients be represented by $(1,...m)$, with associated rates of these combinations of disutility events across patient populations respectively denoted by $(DU_1,...,DU_m)$, and associated values in avoiding each of these combinations of disutility events $(k_1,...,k_m)$. Then, under the first assumption of the correspondence theorem, net benefit for any hospital can be represented relative to a common comparator ($comp$) as:

$$NB_i = k_1 \times (DU_{1comp} - DU_{1i}) + ... + k_m (DU_{mcomp} - DU_{mi}) - (C_i - C_{comp}) =$$

$$- (k_1 \times DU_{1i} + .. + k_m \times DU_{mi} + C_i) + (k_1 \times DU_{1comp} + .. + k_m \times DU_{mcomp} - C_{comp}). \quad (5.13)$$

Without loss of generalization, let $NB_i > NB_j$.

Then in (5.13) all comparator cost and effect terms cancel in comparing two hospitals with the same comparator under the second correspondence theorem assumption.

$$\Leftrightarrow \quad - (k_1 \times DU_{1i} + .. + k_m \times DU_{mi} + C_i) > -(k_1 \times DU_{1j} + .. + k_m \times DU_{mj} + C_j). \quad (5.14)$$

Multiplying both sides of (5.14) by minus 1 the sign changes and we translate from maximising net benefit to minimising net loss per admission.

$$\Leftrightarrow \quad C_i + DU1_i \times k_1 + ... + DUm_i \times k_m < C_j + DU1_j \times k_1 + ... + DUm_j \times k_m. \quad (5.15)$$

QED (multiple events and a common comparator case)

5.3.4 Proof - Case 4: multiple disutility bearing events and differences in expected costs and effects

Let all potential combinations of disutility events observed in patients be represented by $(1,...m)$ with rates of these combinations across patient populations $(DU1,...,DUm)$ and associated values from avoiding these combinations of disutility events be $(k_1,...,k_m)$. Then under the first assumption of the correspondence theorem, net benefit for any hospital can be represented relative to a comparator as:

$$NB_i = k_1 \times (DU1_{comp_i} - DU1_i) + ... + k_m (DUm_{comp_i} - DUm_i) - (C_i - C_{comp_i}) =$$

$$- (k_1 \times DU1_i + .. + k_m \times DUm_i + C_i) + (k_1 \times DU1_{comp_i} + .. + k_m \times DUm_{comp_i} - C_{comp_i}). \quad (5.16)$$

Without loss of generalization (order is arbitrary in establishing a correspondence), let $NB_i > NB_j$.

Then from (5.16) $\Leftrightarrow$

$$- (k_1 \times DU1_i + .. + k_m \times DUm_i + C_i) > -(k_1 \times DU1_j + .. + k_m \times DUm_j + C_j) + z, \quad (5.17)$$

where: $z = k_1 \times (DU1_{comp_j} - DU1_{comp_i}) + .. + k_m \times (DUm_{comp_j} - DUm_{comp_i}) + C_{comp_j} - C_{comp_i}$

Multiplying both sides of (5.17) by minus 1 the sign changes and we translate from maximising net benefit to minimising net loss per admission:

$$\Leftrightarrow \quad C_i + DU1_i \times k_1 + ... + DUm_i \times k_m < C_j + DU1_j \times k_1 + ... + DUm_j \times k_m - z. \qquad (5.18)$$

Therefore, under the second assumption, if differences in expected costs and disutility events are adjusted for, there is a one-to-one correspondence with maximising net benefit.

QED (multiple events and adjusting for difference in expected costs and effects)

The correspondence theorem allows relative performance measurement conditional on the decision maker's value of avoiding disutility events consistent with maximising net benefit per admission, where relative disutility event rates represent effects of care. An ordinal[3] one-to-one correspondence results from a linear translation between maximising net benefit and minimising cost plus disutility events valued as in net benefit.

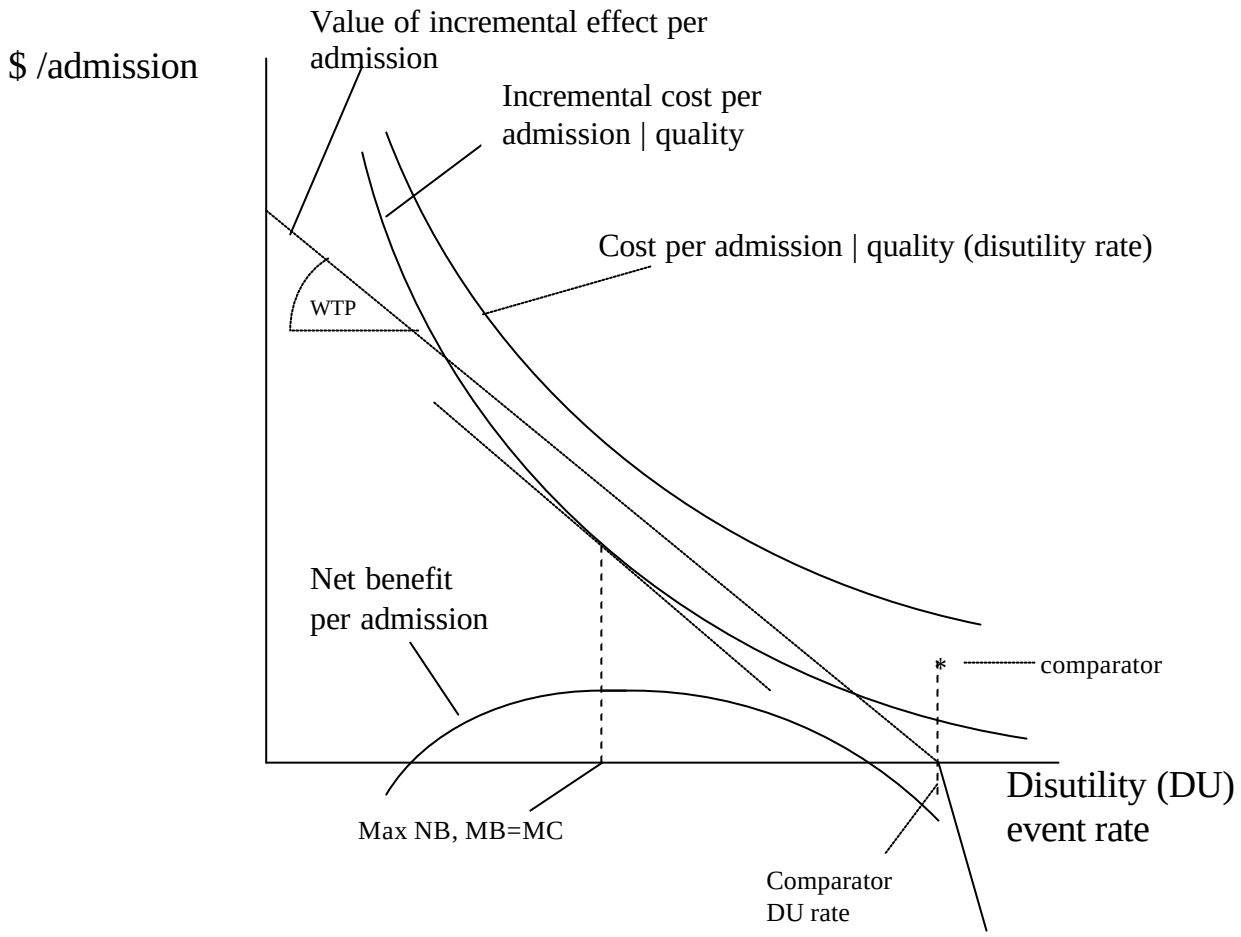
While expected costs and disutility event rates can differ due to differences in populations, access to technology or potentially size of operation, correspondence conditions are able to be feasibly satisfied when applied at a clinical activity (DRG) level. At this level, patient populations are naturally more homogenous, disutility events can be identified to cover effects of care and differences in patient risk factors can be flexibly adjusted for. As described in chapter 3, a DRG level of analysis also allows identification of inefficiency hidden by aggregation, while avoiding the Fox (1999) aggregation paradox and confounding of effects across hospital activities at aggregate levels. Methods of adjustment for differences in expected costs and effects with DEA, including standardization prior to efficiency assessment, restricting comparison sets to equivalent environment, second stage regression and peer grouping, are outlined in detail in section 5.7.2.

[3] The one-to-one correspondence, while proven to be ordinal here, is later shown to also be cardinal in differences in net benefit in section 6.3.1 in identifying funding models with appropriate incentives. For purposes of relative performance measurement in peer identification, only the ordinal correspondence is required, although as section 6.6 illustrates, the cardinal nature of correspondence also allows interpretation of economic, technical, scale and allocative inefficiency, in terms of differences in net monetary or effect benefit at a given WTP.

5.4 Interpreting net benefit maximisation in cost-disutility space

Under correspondence conditions, minimising cost plus disutility events valued as in net benefit provides an appropriate trade-off between the value of effects and costs of quality of care in ordering performance consistent with maximising net benefit per admission. This is illustrated in figure 5.2, in relative comparison of net benefit per admission for hospitals, relative to a common comparator, with effects of care measured by the rate of a disutility event (a lower rate indicating higher quality).

Figure 5.2 Net benefit per admission in cost-disutility space



For a given technology, a frontier representing the minimum cost per admission conditional on disutility bearing event rate is, *a priori*, expected to reflect diminishing returns as disutility event rate approaches 0. In figure 5.2 this is reflected in the incremental cost of reducing the disutility event rate increasing as the rate of disutility events approaches 0.

In considering relative net benefit of hospitals on the frontier, incremental cost per admission represents a parallel shift down, in the vertical plane, of the frontier of minimum cost conditional on quality, by the cost per admission of a common comparator. Importantly for illustrating correspondence, this implies that the slope of the minimum cost curve and incremental cost curve are the same at any disutility event rate. While an optical illusion is created in figure 5.2 that these cost curves move closer together as disutility event rates fall, this is because the eye looks at the shortest distance (radially), rather than in the vertical plane, where the curves are parallel.

In a net benefit framework the comparator may be considered that of next best alternative care. However, as the correspondence theorem proof in section 5.3 demonstrated, where there is a common comparator, the cost and effect of this comparator does not affect correspondence, assuming a constant value of willingness to pay (k) for effects of care.[4] Relative ordering of net benefit per admission and sum of cost per admission and disutility event rate multiplied by a constant value (k) are the same where correspondence assumptions are met.

In considering the incremental value of the effects of care for a hospital relative to a common comparator, this incremental value is 0 at the same disutility event rate as the comparator. Incremental effect becomes positive as the disutility event rate reduces below (quality of care increases above) that of the comparator and negative if the disutility event rate is above (quality of care below) that of the common comparator. In a net benefit framework the value of avoiding disutility events, and hence slope on the

[4] The effect on performance measurement of varying from an assumption of a constant k in net benefit is considered in section 5.7.3.

curve, is characterised as the decision maker's willingness to pay threshold, a constant. This assumes that losses are valued equally to gains for the same health increment. Across a population the value given to a unit of health by decision makers can, however, be characterised as the willingness to pay for positive incremental effect, relative to the expected effect of a comparator representing a minimum endowment or entitlement to care (such as next best alternative form of care), and willingness to accept for negative incremental effect. Under prospect theory (Kahnemann and Tversky, 1979), with loss aversion, willingness to accept health loss is predicted to be valued more than willingness to pay for equivalent health gain. Empirical support for this in health care (Willan, O'Brien and Leyva, 2001), with a 2-3 fold greater value for WTA than WTP for equivalent health gain, suggests a kink[5] in the incremental value of effects of care at the comparators level of effects of care as depicted in figure 5.2. The implications of this kink for relative performance measurement under the correspondence theorem are considered in section 5.7.3. In general, in using WTP the benefit of doubt is given to hospitals with relatively low quality, assuming that they have a disutility event rate below that (quality above that) of an unknown comparator.

From the incremental cost and incremental effect curves the net benefit curve can be constructed, following equation 5.1, as incremental value of effects less incremental cost, at any disutility event rate. Assuming diminishing marginal returns to resources in reducing disutility events, net benefit per admission is maximised at the disutility event rate where marginal cost is equal to marginal benefit. That is, at the disutility event rate in figure 5.2, where the slope of the incremental cost curve, conditional on disutility event rate, is equal to the slope of the incremental effect curve. Assuming that quality of hospital care is above a minimum threshold (endowment) level at net benefit maximisation and a constant value of willingness to pay to avoid disutility events, net benefit maximising quality of care will be where the marginal cost of increasing quality (reducing disutility event rates) equals the willingness to pay to avoid disutility event rates (k in figure 5.2).

[5] O'Brien, Gersten, Willan and Faulkner (2002) suggested a kink in the threshold regions in the incremental cost effectiveness plane in health technology assessment. Section 8.5.2 also considers a kink in isocost curves in comparing relative performance of strategies for health technology assessment in the cost-disutility plane.

A level of quality (disutility event rate) maximising net benefit represents the ideal trade-off between the cost and value of quality of care. Below this level of quality (above this disutility event rate) the marginal cost of increasing quality is less than marginal value, while above this level of quality marginal cost is greater than marginal value. In comparison, cost minimisation occurs where the marginal cost of reducing quality is 0. Cost minimising quality of care will only coincide with net benefit maximising quality of care if inframarginal costs of treating disutility events outweigh intramarginal costs of increasing quality, down to a disutility event rate of 0. That is, if expected costs of care continually monotonically decrease with reduction in disutility event rates.

However, it should be noted that in comparison of relative performance across hospitals, maximising net benefit does not necessarily guarantee net benefit is positive. Whether net benefit is positive is an allocative efficiency question of the appropriateness of technology used in hospital care compared to other forms of care (strategies or technologies), rather than one of relative efficiency across providers in using a given technology. This question is appropriately addressed by health technology assessment and can only be answered if comparator/s and costs and effects of care are identified. Where multiple technologies are compared with health technology assessment, application of the net benefit correspondence theorem with use of frontiers in the cost-disutility plane are illustrated in chapter 8 to have distinct advantages over current frontiers constructed in the incremental cost effectiveness plane.

5.5 Application of the correspondence theorem to relative performance measurement

In applying the correspondence theorem to performance measurement, index and frontier methods can be employed that appropriately minimise the objective function of cost per admission plus disutility events, valued at the decision makers threshold of WTP.

5.5.1 Index methods

The simplest method to apply the correspondence theorem in relative performance measurement at a clinical activity level across inpatient care in public hospitals is to minimise the objective of costs plus disutility events per admission valued at k as in equation (5.6). The reciprocal of this objective function, measured across hospitals, can be used as a simple index of relative performance allowing for health related quality of care consistent with maximising net benefit per admission, under net benefit correspondence theorem assumptions.[6]

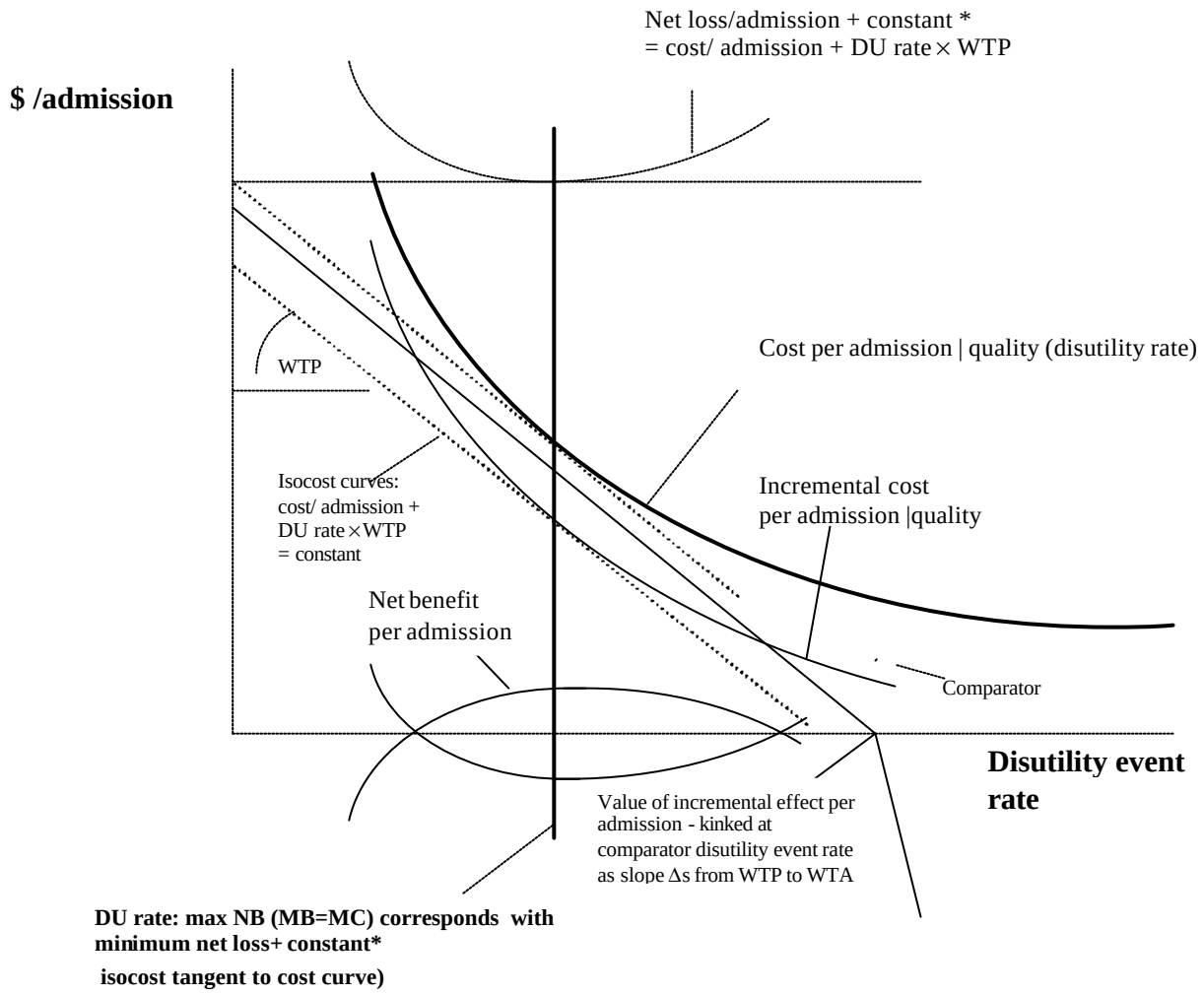
5.5.2 Frontier methods

To allow a richer story in performance measurement, the objective function in equation (5.6) can be minimised using frontier methods. This allow decomposition of economic efficiency consistent with net benefit into sources of technical, allocative and, where appropriate, scale and congestion efficiencies.

Specifying admissions as an output, and costs and disutility events as inputs (priced at their value of being avoided), allows an objective function, with input-orientated economic efficiency of minimising cost per admission plus disutility events, valued as in net benefit. Under constant returns to scale, hospitals are compared against the same reference peer, regardless of size, and hence the ordering of economic efficiency with this specification will correspond to that required in equation (5.6). Graphically, in figure 5.3, isocost curves of the form, cost plus disutility events valued at WTP (k), are minimised at their point of tangency to the unit isoquant representing minimum combinations of cost and disutility events per admission.

[6] Applying index methods to measure performance over time under the correspondence theorem is considered in section 7.4.2

Figure 5.3 Correspondence between maximising net benefit per admission and minimising costs plus disutility events valued at WTP



* the constant is equal to the cost per admission for the comparator plus the value of the comparators disutility rate, a positive constant assuming hospitals face the same comparator

Under the correspondence theorem assumptions, maximising relative performance at this point of tangency represents an appropriate trade-off between the dual objectives for hospitals, suggested by Harris (1977), of administrators minimising costs and clinicians maximising health effects of care.

For each hospital, economic efficiency is measured in radial contraction of costs and disutility events per admission to the isocost curve, with a slope of WTP tangent to the best practice frontier. For any given willingness to pay (k in equation 5.1), the point of tangency between an isocost curve with slope k and a convex best practice frontier minimising costs and disutility events per admission, minimises the value of isocost curves in the feasible set, and hence the objective function, in equation (5.6).

The ordering of net benefit has a correspondence not just on the frontier, but in any bilateral comparison, as shown in the proof of the correspondence theorem in section 5.3. For a given WTP, isocost curves can therefore be drawn for any hospitals through their observed combination of cost and disutility bearing event rate, to determine a complete ordering of relative performance consistent with maximising net benefit under correspondence conditions.

Under correspondence conditions, cost per admission plus value of disutility event rate is a vertical displacement of the net loss function, which in turn is a negative transformation of the net benefit function to be maximised, as in the correspondence theorem proof. As these are linear translations in the vertical plane, there is a one-to-one correspondence between maximising net benefit per admission, and minimising net loss, plus a constant, per admission.

Considering the correspondence in marginal terms, as the incremental cost curve is a vertical displacement of the cost curve for a common comparator, these two curves have the same marginal cost (slope) for any given disutility event rate (quality of care). Therefore, at the disutility event rate where net benefit is maximised and marginal cost is equal to marginal benefit, an isocost curve with slope of marginal benefit must be tangent

to the minimum cost curve (slope marginal cost). Given a constant value of avoiding disutility events and diminishing returns to resources in increasing quality of care, marginal cost of quality is less than marginal value of quality above this disutility event rate. Conversely, below this disutility event rate the marginal cost of quality is greater than marginal value of quality.

5.5.3 Interpreting economic efficiency minimising cost & value of disutility events

As argued in chapter 2, to allow a trade-off between public hospital objectives of cost minimisation and health maximisation and to create appropriate incentives for quality of care, the production frontier underlying efficiency measurement (technical, allocative or economic) needs to reflect the value of effects relative to the resource use of quality of care. Analogous to profit maximisation, maximising net benefit (Stinnett and Mullahy, 1998) per admission as the incremental value of effects of care less the cost of care per admission, explicitly allows such a trade-off.

Provided assumptions of the correspondence theorem are satisfied, economic efficiency, measured in minimising the sum of cost plus disutility events valued as in net benefit may be better described as net benefit efficiency of public hospitals. Under net benefit correspondence theorem assumptions, this economic efficiency measure provides the same relative ordering of hospitals as maximising net benefit. Relative performance is measured as improving while the marginal benefit of increasing quality is greater than the marginal cost of increasing quality.

5.5.4 Technical efficiency minimising cost and disutility events per admission

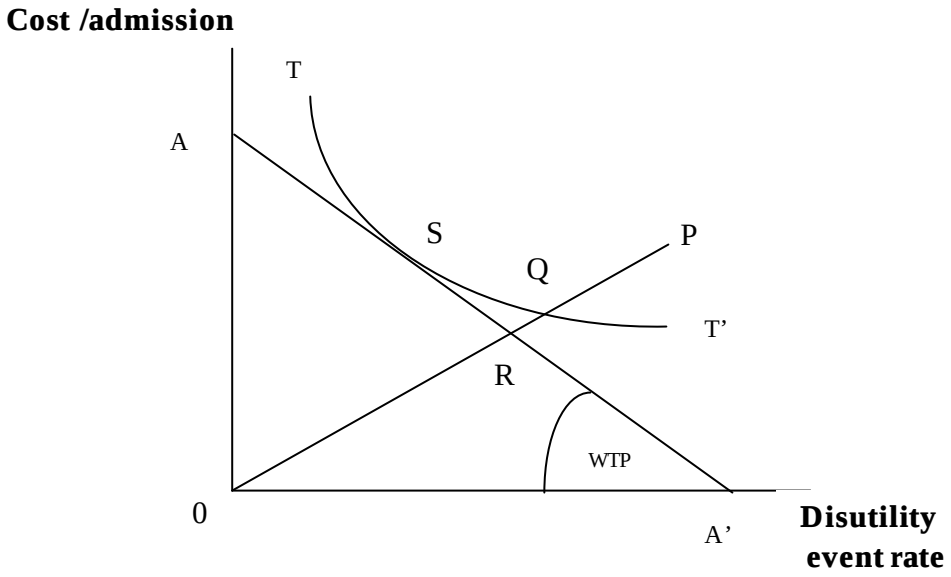
Frontier methods allow estimation of input-orientated technical efficiency under the Debreu (1951) and Farrell (1957) definition, of one minus the maximum equi-proportionate reduction of all inputs that still lie in the feasible set, allowing production of given outputs. In decomposing economic efficiency from an input-orientation, technical efficiency reflects the ability of a firm to minimise inputs for a given level of outputs, and allocative efficiency the ability of a firm to use inputs in the correct proportions, given their relative prices and production technology. Using distance

functions (Shephard, 1953; Farrell, 1957) with an input-orientation, technical efficiency can be estimated as a radial contraction to the best practice frontier, and allocative efficiency as a residual of economic and technical efficiency, as described in appendix 4.1.

The correspondence theorem demonstrated that minimising cost plus disutility events, valued at the decision maker's threshold value of WTP, orders performance consistent with net benefit maximisation, where disutility events represent effects of care and hospitals face a common comparator. Under correspondence conditions, net benefit per admission is maximised where isocost curves reflecting relative value of disutility events (slope of WTP) are tangent to the unit isoquant, with cost and disutility events as inputs and admissions as outputs, as depicted in figure 5.3.

Without considering effects of quality of care, technical or cost efficiency in production of admissions is neither sufficient, nor necessary, for efficiency in production of health or net benefit maximisation, in treating patient populations. This is clearly illustrated by the frontier minimising cost and disutility events representing quality of care. Cost efficiency in production of admissions *per se*, at cost minimising point A in figure 5.1, is not necessary for net benefit maximisation (at B). However, being on the frontier minimising cost conditional on disutility events, is a necessary (while not sufficient) condition for net benefit maximisation, under correspondence conditions.

Figure 5.4: Economic, technical and allocative efficiency of net benefit



In figure 5.4, radial contraction to the unit isoquant TT' can be described under correspondence conditions as technical efficiency in production of net benefit. Such technical efficiency is necessary, while not sufficient, for net benefit maximisation. The joint maximand that this technical efficiency measure represents in minimising costs and disutility events per admission is intuitively appealing. It reflects the twin objectives of hospitals in treating patients, characterised by Harris (1977) as maximising health gain (medical staff) and minimising costs (administrators).

For a hospital at P in figure 5.4, technical efficiency (of net benefit) can be estimated as OQ/OP, representing radial contraction to the frontier TT', which minimises cost and disutility events per admission. The residual of economic efficiency in maximising net benefit and technical efficiency in minimising cost and disutility events (OR/OQ in figure 5.4) can be described as allocative efficiency in net benefit maximisation. It reflects loss of net benefit from hospitals failing to appropriately trade-off the value of clinical quality of care (reduction in disutility event rates) against cost of quality of care, given the decision maker's value of effects of care.

As there is an assumption that minimising costs and disutility bearing events is an appropriate objective for hospitals, what has been described here as technical efficiency of net benefit production, may be strictly considered a measure of economic efficiency. This is particularly as costs, rather than input data, have been used, under the assumption that hospitals face the same factor prices, an assumption also made with the current measurement of cost per case-mix adjusted separation. If input data were available, what may be considered as true technical efficiency of net benefit production under correspondence, in minimising disutility events and inputs per admission, could be calculated. Allocative efficiency would then reflect the ability to use inputs in the correct proportions, given their prices and value of quality (avoiding disutility event rates) in production of net benefit.

However, in the absence of data on physical inputs at a DRG level, I have described the radial contraction to a frontier minimising cost and disutility event per admission, under correspondence conditions, as a measure of technical efficiency in the production of net benefit throughout this thesis. The residual of net benefit efficiency (under correspondence conditions) and this technical efficiency measure I refer to as a measure of allocative efficiency in production of net benefit.

5.6 Illustrating relative performance measurement applying the correspondence theorem

Application of the correspondence theorem to relative performance measurement is illustrated with routinely collected cost, activity (admissions) and effect (mortality) data for an individual clinical activity (DRG E62a). As in chapter 4, relative performance of 45 Australian public hospitals for DRG E62a (respiratory infection) is compared using data envelopment analysis (DEA). However, unlike analysis in chapter 4, analysis here is input-orientated, with disutility event rates (mortality) and costs specified as inputs, and admissions as an output.

For purposes of illustrating application of the correspondence theorem, hospitals are assumed to face the same expected costs and disutility event rates given patient

populations treated within this DRG, and relative mortality rates across hospitals for this DRG are assumed to cover effects of care with no difference in incremental rates beyond separation. In reality, lack of adjustment for patient risk factors, and the lack of data linkage to effects beyond separation are limitations in this illustration. To robustly apply the correspondence theorem, relative performance measurement should adjust for pre-admission, and post-separation, differences, to overcome potential biases from lack of homogeneity in patient populations and admissions as a representation of care, and prevent incentives for cream-skimming and cost-shifting respectively.

In applying the correspondence theorem to performance measurement, the limitation in this thesis in adjusting for patient risk factors and post-care effects is, however, one of data available, rather than one of the method or ability to identify and obtain data in general. As section 5.7 addresses in detail, the linear nature of correspondence allows inclusion of multiple effects, including those beyond point of separation. At a clinical activity level, patient risk factors can be adjusted for, and decision analytic methods employed to systematically identify effects of quality of care, including those requiring data linkage and perceived utility bearing aspects of quality framed as disutility events. Correspondence theorem conditions can, therefore, be comprehensively and robustly satisfied at a clinical activity level, employing decision analytic methods, as in health technology assessment. The net benefit correspondence theorem framework allows policy makers to focus on what is required to avoid exogenous biasing factors pre- and post-care in performance measurement.

While data was not available to adjust for patient risk factors and post-care effects in this thesis, the framework provided by net benefit correspondence theorem conditions appropriately requires that assumptions such as no difference in population risk at admission and coverage and effects beyond care are stated. This explicit nature of this framework has clear advantages over current performance measures such as cost per case-mix adjusted separation. Perverse incentives for cream skimming and cost-shifting exist with cost per case-mix adjusted separation, where within-DRG patient risk factors, and post care effects, are not adjusted for in measuring relative performance. However,

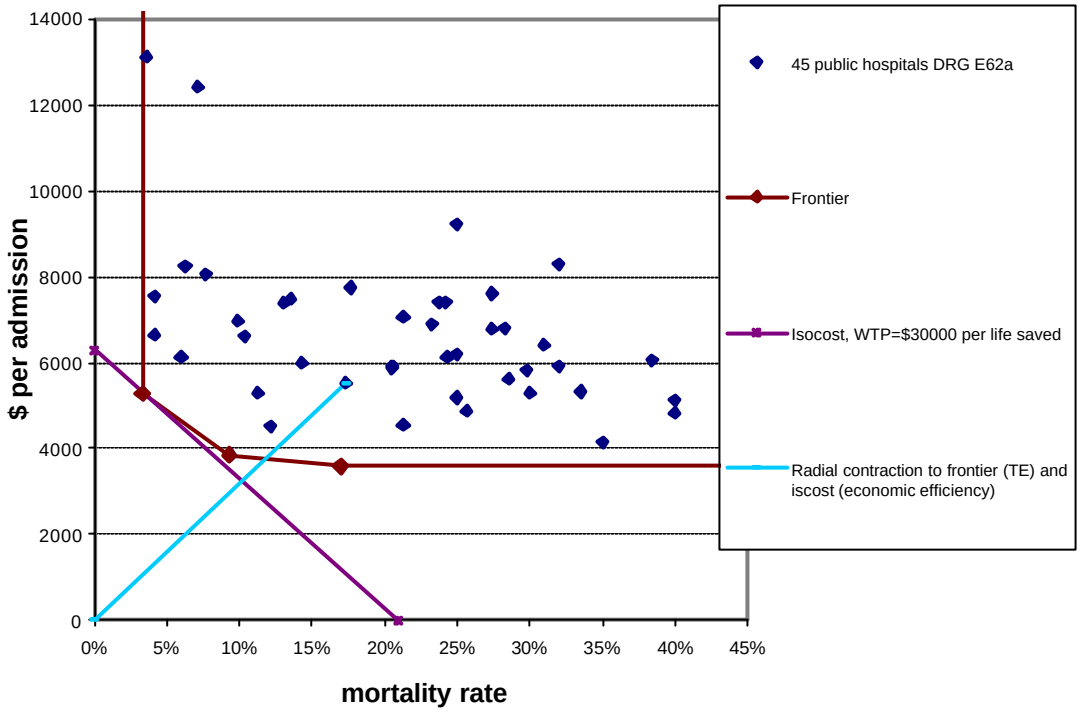
in comparing estimates of cost per case-mix adjusted admission, no conditions are required to be satisfied. Implicit assumptions are therefore ignored such as: no differences in quality or effects of care; homogeneity of within-DRG risk factors and; no difference in cost or effects beyond-care. The absence of requirements to state these assumptions reflects the lack of an appropriate objective function, or theoretical basis, for measuring performance with cost per case-mix adjusted separation. It also reflects the inability at an aggregate level, to flexibly identify or synthesise quality of care, patient level risk factors or effects beyond-separation.

The correspondence theorem provides a method for net benefit maximisation, but also challenges policy makers to meet existing policy challenges for risk adjustment and data linkage, to allow robust performance measurement and funding. Using decision analytic methods at a clinical activity level are shown, in section 5.7 and chapter 7, to focus data linkage and risk adjustment, allowing correspondence conditions to be comprehensively and systematically satisfied. Meeting these policy challenges allows evidence-based performance measurement, providing incentives for net benefit maximising quality of care, while avoiding incentives for cream-skimming and cost-shifting.

5.6.1 Economic and technical efficiency of net benefit

Figure 5.5 illustrates technical and economic efficiency, of net benefit under correspondence assumptions, estimated for 45 hospitals at a clinical activity (DRG E62a), relative to a frontier minimising cost and disutility event (mortality), per admission.

Figure 5.5: Relative performance of 45 Australian hospitals for DRG E62a in minimisng cost and mortality per admission (technical efficiency) and maximising net benefit at a decision maker's willingness to pay per life saved (economic efficiency)



For each hospital, cost per admission and disutility events per admission are equi-proportionally reduced (radial contraction). Technical efficiency, under constant returns to scale, is estimated using distance functions (Farrell, 1957; Shephard, 1970) with radial contraction to the frontier of best practice, while economic efficiency is estimated with radial contraction to an isocost line (with slope of negative WTP[7]), tangent to this frontier. Economic efficiency, corresponding to net benefit maximisation under correspondence conditions, is dependent on the decision-maker's value of avoiding disutility events, while technical efficiency (of net benefit) is not.

Where uncertainty exists as to the value of decision-makers' threshold, relative economic efficiency can be conditioned on potential values. While a willingness to pay (k) of $30,000 per life saved is used illustratively in figure 5.5, table 5.6.1 presents the relative economic efficiency at values of 0, $5,000, $10,000, $25,000, $50,000 and $100,000 per life saved.

[7] The slope represents the decision maker's monetary value of willingness to pay to avoid disutility events (k in equation 5.1).

Table 5.6.1 Relative hospital performance (economic efficiency) for DRG E62a, conditional on value of avoiding mortality

Hospital	$0	$5,000	$10,000	$25,000	$50,000	$100,000
1	0.74	0.63	0.54	0.41	0.28	0.19
2	0.39	0.41	0.41	0.40	0.32	0.25
3	0.45	0.51	0.54	0.61	0.58	0.55
4	0.29	0.34	0.37	0.43	0.43	0.44
5	0.70	0.61	0.53	0.40	0.28	0.19
6	0.44	0.51	0.54	0.62	0.61	0.59
7	0.87	0.73	0.63	0.47	0.32	0.22
8	0.60	0.64	0.65	0.64	0.53	0.42
9	0.49	0.54	0.55	0.57	0.50	0.42
10	0.54	0.63	0.68	0.80	0.80	0.80
11	0.48	0.56	0.6	0.71	0.72	0.74
12	0.43	0.44	0.42	0.38	0.29	0.21
13	0.59	0.54	0.48	0.39	0.27	0.19
14	0.27	0.33	0.36	0.44	0.47	0.52
15	0.54	0.61	0.63	0.66	0.59	0.51
16	0.58	0.58	0.55	0.49	0.37	0.28
17	**0.93**	**1.00**	**1.00**	**0.99**	**0.81**	**0.65**
18	0.48	0.50	0.49	0.45	0.36	0.27
19	0.79	0.84	0.84	0.81	0.66	0.52
20	0.59	0.59	0.56	0.5	0.38	0.28
21	0.48	0.53	0.54	0.56	0.49	0.41
22	0.74	0.70	0.64	0.54	0.39	0.28
23	0.61	0.63	0.6	0.56	0.43	0.33
24	0.68	0.64	0.58	0.48	0.34	0.24
25	0.79	0.77	0.72	0.62	0.46	0.33
26	**1.00**	**0.97**	**0.91**	**0.78**	**0.58**	**0.42**
27	0.59	0.67	0.71	0.80	0.76	0.71
28	0.46	0.50	0.50	0.50	0.43	0.34
29	0.68	0.74	0.75	0.75	0.64	0.52
30	0.61	0.58	0.53	0.44	0.32	0.23
31	0.65	0.68	0.66	0.62	0.49	0.38
32	0.53	0.53	0.50	0.45	0.34	0.25
33	**0.68**	**0.79**	**0.85**	**1.00**	**1.00**	**1.00**
34	0.51	0.58	0.60	0.65	0.58	0.51
35	0.48	0.50	0.49	0.46	0.36	0.28
36	0.69	0.67	0.62	0.53	0.39	0.29
37	0.62	0.59	0.54	0.46	0.34	0.24
38	0.52	0.54	0.52	0.48	0.38	0.29
39	0.56	0.54	0.50	0.43	0.32	0.23
40	0.61	0.62	0.60	0.55	0.43	0.33
41	0.64	0.61	0.57	0.48	0.35	0.25
42	0.51	0.53	0.52	0.49	0.39	0.3
43	0.67	0.62	0.55	0.45	0.31	0.22
44	0.47	0.48	0.46	0.42	0.33	0.25
45	0.53	0.53	0.50	0.44	0.33	0.25

In table 5.6.1, net benefit maximising peers (bolded with economic efficiency of 1), benchmarks (cost and disutility event rates of peers) and relative ordering, change across potential ranges for value of quality (WTP to avoid disutility events). Including a value of avoiding disutility bearing events of 0 allows comparison with hospital performance minimising cost per admission. At a value of $0 per life saved, hospital 26 is a peer and benchmark with the lowest cost per admission of $3590 per admission, while hospital 33 with cost per admission of $5283 per admission has economic efficiency of 0.68. In comparison at $50,000 per life saved, hospital 33 is the benchmark with 3.0% mortality and cost of $5283 per admission, while hospital 26, with 17% mortality and cost per admission of $3590, has economic (net benefit) efficiency of 0.58.

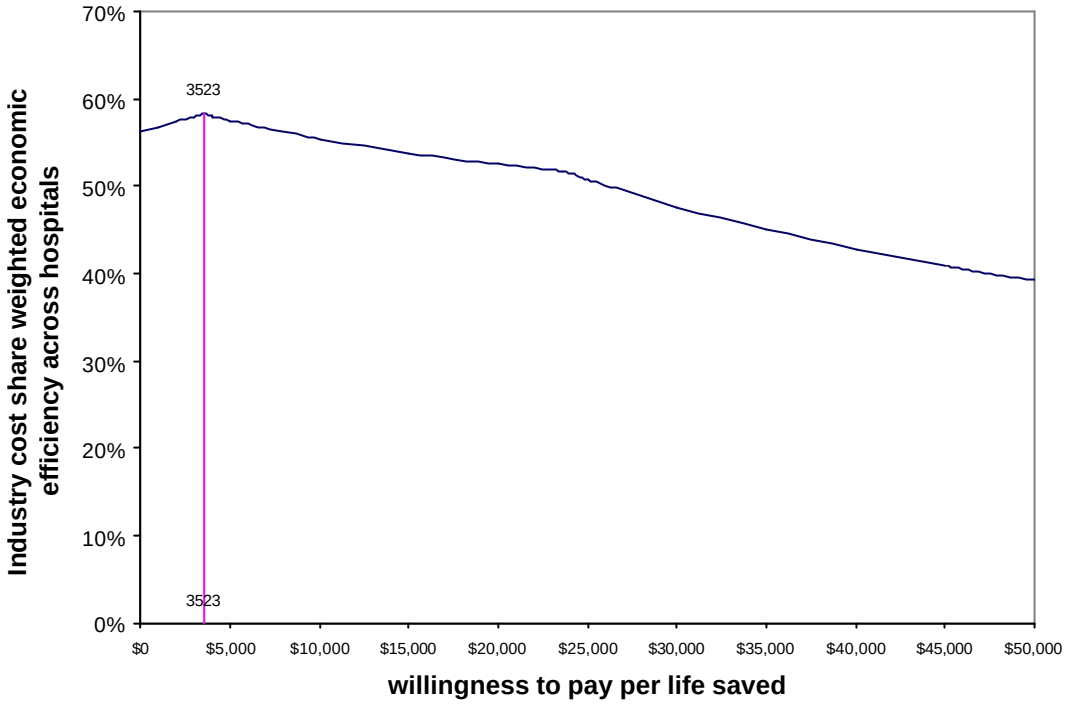
Where correspondence theorem assumptions are satisfied, relative performance measured consistent with maximising net benefit at the decision maker's threshold value (k=WTP), can be compared with that of cost minimisation (k=0). The degree of divergence, between relative economic efficiency measured at WTP and a 0 value, indicates the level of inappropriateness in measuring performance with cost per admission, implicitly including the cost but ignoring the value of quality of care. Measuring relative performance and identifying peers clearly does not create incentives for net benefit maximising quality of care. Cost minimisation conditional on health related quality of care, rather than cost minimisation *per se*, is the necessary, while not sufficient, condition for net benefit maximisation in efficient provision of health care.

5.6.2 Implicit industry value of quality (shadow price)

Conditioning on the value of avoiding disutility bearing events is equivalent to rotating isocost curves around the best practice frontier, under the assumption of constant returns to scale. Identifying industry economic efficiency as the cost-share-weighted average of individual hospitals, the implicit industry value of quality, in current industry behavior,

can be estimated as the WTP that maximises industry economic efficiency. The WTP where industry economic efficiency is maximised will also correspond to where allocative efficiency of net benefit (as the residual ratio of technical and economic efficiency of net benefit) is maximised. In our example, economic efficiency of the industry is maximised where the relative price or value of avoiding death is $3523 per death avoided, as illustrated in figure 5.6.

Figure 5.6: Industry value of quality of care for DRG E62a (cost per life saved)



5.6.3 Technical efficiency under CRS, VRS and NIRS and scale efficiency

Technical efficiency of net benefit under constant returns to scale (CRS) can be estimated relative to the unit isoquant minimising costs and disutility events per admission using the CRS form of DEA (Charnes, Cooper and Rhodes, 1978). To explore questions of scale effects, technical efficiency under constant returns to scale can be further decomposed into technical efficiency under variable returns to scale and scale efficiency as a residual (Färe, Grosskopf and Logan, 1983; Banker, Charnes and Cooper, 1984),. This decomposition is not dependent on the willingness to pay threshold in avoiding disutility events. A non-increasing returns to scale formulation of DEA [Coelli (1998:151), Färe, Grosskopf and Lovell (1994:50)], can also be used to identify whether scale inefficiency is due to increasing returns to scale (IRS) or decreasing returns to scale (DRS). The methods for each of these decompositions are described in appendix 4.1. Estimates of technical efficiency, under CRS and decomposition to estimate scale efficiency, are presented for each of the 45 hospitals in table 5.6.2.

Table 5.6.2 Technical efficiency under constant, variable and non-increasing returns to scale and scale efficiency

Hospital	Technical efficiency (constant returns to scale)	Technical efficiency (variable returns to scale)	Scale efficiency	Technical efficiency (NIRS)*
1	0.74	1.00	0.74	IRS
2	0.41	0.74	0.56	IRS
3	0.61	1.00	0.61	IRS
4	0.47	1.00	0.47	IRS
5	0.70	0.84	0.83	IRS
6	0.62	1.00	0.62	IRS
7	0.87	0.98	0.88	IRS
8	0.65	0.82	0.79	IRS
9	0.58	0.68	0.86	IRS
10	0.80	1.00	0.80	IRS
11	0.80	1.00	0.80	IRS
12	0.44	0.48	0.93	IRS
13	0.59	0.64	0.92	IRS
14	0.93	1.00	0.93	IRS
15	0.67	0.73	0.92	IRS
16	0.59	0.62	0.96	IRS
17	**1.00**	**1.00**	**1.00**	
18	0.51	0.52	0.98	IRS
19	0.847	0.849	0.998	IRS
20	0.60	0.61	0.98	IRS
21	0.57	0.57	0.99	IRS
22	0.74	0.76	0.97	IRS
23	0.633	0.634	0.999	IRS
24	0.68	0.70	0.97	IRS
25	0.79	0.80	0.99	IRS
26	**1.00**	**1.00**	**1.00**	
27	0.80	0.81	0.99	DRS
28	0.51	0.58	0.88	DRS
29	0.76	0.87	0.88	DRS
30	0.61	0.74	0.82	DRS
31	0.68	0.84	0.82	DRS
32	0.54	0.70	0.77	DRS
33	**1.00**	**1.00**	**1.00**	
34	0.65	0.75	0.87	DRS
35	0.51	0.70	0.73	DRS
36	0.69	0.98	0.71	DRS
37	0.62	0.88	0.70	DRS
38	0.54	0.79	0.69	DRS
39	0.56	0.83	0.68	DRS
40	0.63	1.00	0.63	DRS
41	0.64	0.97	0.66	DRS
42	0.54	0.98	0.55	DRS
43	0.67	1.00	0.67	DRS
44	0.49	1.00	0.49	DRS
45	0.54	1.00	0.54	DRS

* scale inefficiency due to increasing returns to scale (IRS) or decreasing returns to scale (DRS)

It should be stressed that in estimating technical efficiency under variable returns to scale, and consequently scale efficiency, the assumption of best practice representing technology in DEA is effectively required across all scales of production, as described in appendix 4.1. Variable returns to scale (VRS), or non-increasing returns to scale (NIRS), formulations of DEA require enough decision making units (DMUs) at different scales of production, for observed best practice to robustly reflect technology. With 45 hospitals, estimates of technical efficiency under VRS and scale efficiency, in table 5.6.2, are unlikely to allow robust estimation of VRS technology across all scales of production. Measured scale inefficiency is, therefore, likely to be upwardly biased for hospitals and should be interpreted illustratively, rather than prescriptively.

5.6.4 Best practice regions for potential peers

Hospitals, on the technical efficiency frontier in the cost-disutility plane represent all hospitals that can be potentially economically efficient at a given WTP. Regions of WTP, over which technically efficient hospitals (minimising cost and disutility events per admission), are economically efficient, can be simply identified under CRS, by back-solving for adjacent technically efficient hospitals (i,j) on the frontier using:

$$C_i / y_i + DU_i / y_i \times k = C_j / y_j + DU_j / y_j \times k$$
$$\Leftrightarrow k = (C_j / y_j - C_i / y_i) / (DU_i / y_i - DU_j / y_j) \qquad (5.10)$$

Where

C_i, y_i and DU_i are costs, admissions and disutility events for hospital i, respectively.

Illustrating with our case example, hospitals 26, 17 and 33 lie on the frontier, have technical efficiency of 1, and hence each maximise net benefit at some WTP. The cost and mortality rates per admission of these technically efficient hospitals are presented in table 5.6.3.

Table 5.6.3: Cost per admission and mortality rate for potential peers

	Cost/admission	Mortality rate
hospital 26	$3,590	17.0%
hospital 17	$3,858	9.4%
hospital 33	$5,283	3.3%

Ordering these technically efficient hospitals by cost per admission (to allow adjacency on the frontier) and back-solving with (5.10), estimated regions of WTP over which these hospitals are economically efficient are:

hospital 26 from $0 to $3,523 per death avoided;

hospital 17 from $3,524 to $24,356 per death avoided and;

hospital 33 for $24,357 or more per death avoided.

In general, decision makers can be informed of appropriate peers, for any potential threshold value, in avoiding disutility events and, implicitly, value of quality of care, in back-solving for threshold value between adjacent technically efficient hospitals. Where there are n types of disutility events, regions of best practice will be n-dimensional equations or regions in space. For example, for two types of disutility events, say (DU_1, DU_2), associated regions of willingness to pay (k_1, k_2) where a technically efficient hospital has best practice, can be found for each technically efficient hospital (i) by solving:

$$C_i / y_i + DU_{1i} / y_i \times k_1 + DU_{2i} / y_i \times k_2 \geq C_j / y_j + DU_{1j} / y_j \times k_1 + DU_{2j} / y_j \times k_2$$
$$\Leftrightarrow k_1 \geq [k_2(DU_{2j} / y_j - DU_{2i} / y_i) + C_j / y_j - C_i / y_i] / (DU_{1i} / y_i - DU_{1j} / y_j)$$
$$\forall j, j \neq i,$$
$$k_1 > 0, k_2 > 0$$

5.6.5 Congestion efficiency

Strong (costless) disposability of inputs and outputs was assumed in estimating technical efficiency in section 5.6.2. Technical efficiency can also, however, be calculated with DEA under the assumptions of weak disposability of inputs and/or outputs. Weak

disposability allows a backward bending portion of the frontier, relative to which technical efficiency can be estimated.

In the hyperbolic approach of Färe, Grosskopf, Lovell and Parsuka (1989), outlined in section 4.7, weak disposability of undesirable outputs, such as pollution, is assumed. Similarly, technical efficiency can be estimated under weak disposability of disutility bearing events specified as inputs.

Färe, Grosskopf and Lovell (1985, 1994) describe the residual of technical efficiency under strong and weak disposability of inputs as congestion efficiency, corresponding to where there is negative marginal product of inputs, as described in appendix 4.1. In justifying the estimation of input congestion in practice, as Coelli (1998:172) states:

> "It is usually argued that the excess input use is due to constraints which are not under the control of the firm."

In hospitals such congestion efficiency could be justified if disutility events were determined as outside the hospital's control. Congestion efficiency, under variable returns to scale, is estimated, in table 5.6.4, as the residual of technical efficiency under variable returns to scale with disutility events (mortality) as a strongly disposable input, and technical efficiency under variable returns to scale, with disutility events (mortality) as a weakly disposable input.

Table 5.6.4 Congestion efficiency as a residual of technical efficiency, with strong and weak disposability of mortality under variable returns to scale

Hospital	Technical efficiency under VRS with mortality strongly disposable	Technical efficiency under VRS with mortality weakly disposable	Congestion efficiency Under VRS
1	1.000	1.000	1.000
2	0.743	0.743	1.000
3	1.000	1.000	1.000
4	1.000	1.000	1.000
5	0.843	0.849	0.994
6	1.000	1.000	1.000
7	0.983	1.000	0.983
8	0.816	0.816	1.000
9	0.679	0.679	1.000
10	1.000	1.000	1.000
11	1.000	1.000	1.000
12	0.478	0.478	1.000
13	0.643	0.646	0.995
14	1.000	1.000	1.000
15	0.728	0.728	1.000
16	0.616	0.616	1.000
17	1.000	1.000	1.000
18	0.515	0.515	1.000
19	0.849	0.849	1.000
20	0.611	0.611	1.000
21	0.575	0.575	1.000
22	0.757	0.758	0.999
23	0.634	0.634	1.000
24	0.696	0.710	0.979
25	0.799	0.799	1.000
26	1.000	1.000	1.000
27	0.812	0.812	1.000
28	0.579	0.579	1.000
29	0.868	0.868	1.000
30	0.740	0.740	1.000
31	0.836	0.836	1.000
32	0.700	0.700	1.000
33	1.000	1.000	1.000
34	0.746	0.746	1.000
35	0.696	0.696	1.000
36	0.977	0.977	1.000
37	0.878	0.878	1.000
38	0.785	0.785	1.000
39	0.825	0.825	1.000
40	1.000	1.000	1.000
41	0.968	0.968	1.000
42	0.979	0.979	1.000
43	1.000	1.000	1.000
44	1.000	1.000	1.000
45	1.000	1.000	1.000

Modelling disutility events as a weakly disposable input allows for the potential of increasing costs per admission, where disutility event rates are above a cost minimising rate. Table 5.6.4 suggests that there is very little evidence of congestion inefficiency under variable returns to scale. Only hospitals 5, 7, 13, 22 and 24 have any difference in technical efficiency under weak, compared to strong, disposability of disutility events (mortality), and the greatest estimate congestion inefficiency is only 2.1% (hospital 24).

The lack of support for congestion inefficiency, under variable returns to scale, could be an artifice of the lack of support for variable returns to scale technology. Congestion inefficiency may be misinterpreted under VRS formulation as scale inefficiency, where high rates of disutility events and cost per admission are associated with small scale of production. However, the lack of congestion efficiency estimated under constant returns to scale in table 5.6.5 indicates that VRS is not hiding congestion efficiency to any substantial degree.

Table 5.6.5 Congestion efficiency as a residual of technical efficiency, with strong and weak disposability of mortality under constant returns to scale

Hospital	Technical efficiency CRS with mortality strongly disposable	Technical efficiency CRS with mortality weakly disposable	Congestion efficiency CRS
1	0.743	0.851	0.874
2	0.415	0.415	1.000
3	0.615	0.615	1.000
4	0.467	0.467	1.000
5	0.701	0.787	0.891
6	0.624	0.624	1.000
7	0.868	1.000	0.868
8	0.648	0.648	1.000
9	0.580	0.580	1.000
10	0.800	0.800	1.000
11	0.800	0.800	1.000
12	0.444	0.444	1.000
13	0.593	0.630	0.942
14	0.933	0.933	1.000
15	0.670	0.670	1.000
16	0.592	0.592	1.000
17	1.000	1.000	1.000
18	0.507	0.507	1.000
19	0.847	0.847	1.000
20	0.599	0.599	1.000
21	0.569	0.569	1.000
22	0.736	0.750	0.981
23	0.633	0.633	1.000
24	0.678	0.701	0.967
25	0.791	0.791	1.000
26	1.000	1.000	1.000
27	0.805	0.805	1.000
28	0.510	0.510	1.000
29	0.763	0.763	1.000
30	0.606	0.621	0.976
31	0.683	0.683	1.000
32	0.541	0.541	1.000
33	1.000	1.000	1.000
34	0.652	0.652	1.000
35	0.508	0.508	1.000
36	0.692	0.694	0.997
37	0.617	0.626	0.986
38	0.543	0.543	1.000
39	0.559	0.561	0.996
40	0.630	0.630	1.000
41	0.639	0.647	0.987
42	0.537	0.537	1.000
43	0.674	0.714	0.944
44	0.489	0.489	1.000
45	0.538	0.538	1.000

The extent of congestion inefficiency under constant returns to scale is suggested to only potentially be significant in small hospitals 1, 5 and 7, which may alternatively be explained by scale inefficiency. In the absence of enough hospitals to robustly support a VRS technology, it is not possible to distinguish whether inefficiency in these hospitals may be due to congestion, or to scale.

Even where congestion efficiency is 'found', if disutility events represent quality of care, congestion inefficiency can be seen as representing an allocative decision in choice of quality of care. If this were the case, the residual of technical efficiency under strong and weak disposability, described as congestion efficiency, is perhaps more appropriately attributed to allocative inefficiency. However, if disutility events are considered as determined outside the hospital's control (and hence not a quality of care indicator, but rather environmental effect) then congestion inefficiency is appropriately measured as separate from technical or allocative inefficiency.

5.6.6 Peer grouping allowing for *a priori* differences

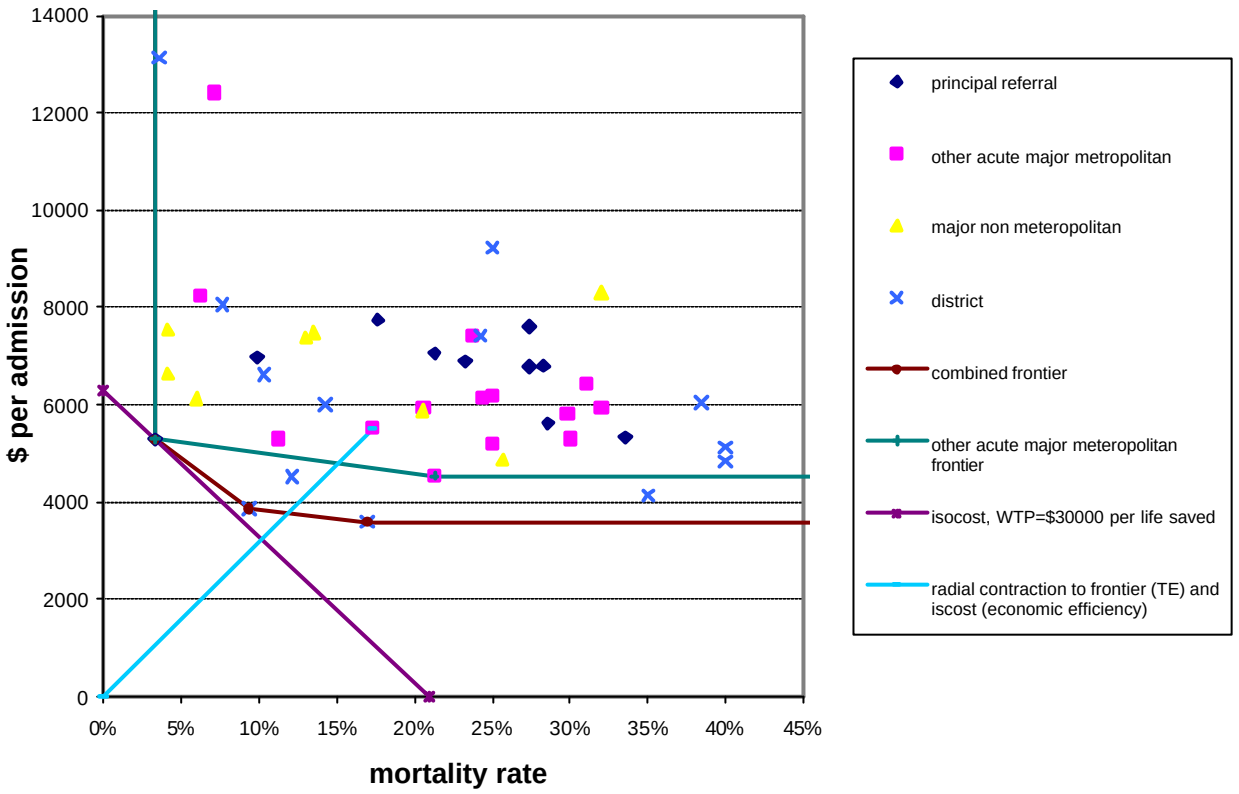
Depending on diagnostic related group, principal referral hospitals in Australia may be expected to have greater patient severity than other acute public hospitals, given referral of more complex cases. In comparison of cost per case-mix adjusted separation in Australia (CDHAC, 2000), such expected differences by hospital type are allowed for by peer grouping. Peer grouping allows for systematic differences in expected costs for the same DRG (due to functions such as teaching, or within DRG patient severity) by comparing relative performance within but not between broad types of hospitals, such as principal referral, other acute, major non-metropolitan and district.

A similar peer grouping approach could also be used in relative performance measurement using frontier methods, and potentially improved upon, if hospital type can be ordered by severity of patients presenting for care (environment). Ordinal environmental factors can be used in identifying appropriate peers for frontier comparison, as proposed by Banker and Morey (1986a), without being as restrictive as

pure peer grouping. Hospitals of a given type (principal referral, other acute, major non-metropolitan and district) can be compared for a given clinical activity with a frontier representing hospital types that, *a priori*, face an equivalent, or more severe patient environment. This approach appropriately increases the reference set from that of pure peer grouping by hospital type.

For example, the referral of the most complex activities, even at a DRG level, to principal referral (public teaching) hospitals in Australia, suggests separate peer grouping, only comparing these hospitals with each other. However, these referral hospitals could still be used in the comparison sets to determine the frontier of best practice for other hospital types. For example, in considering other acute major metropolitan hospitals, they may be *a priori* assumed to have greater severity than all except principal referral hospitals. The performance of principal referral hospitals can, therefore, be measured relative to a frontier constructed including principal referral hospitals, as illustrated in figure 5.7.

Figure 5.7: A frontier for other major acute metropolitan hospitals based on *a priori* ordering on patient severity, principal referral hospitals assumed more severe



The approach of Banker and Morey (1986a) can improve comparator sets relative to that of strict peer grouping. However, the approach does not adjust for differences in risk factors across hospitals within a peer grouping and, as Coelli, Rao and Batesse (1998:167) note, can only deal with one ordinal factor. The assumption of *a priori* ordering, therefore, implicitly assumes homogeneity of relative case-mix, across hospitals within a given type and does not prevent incentives for cream-skimming.

Where there are no reasons for *a priori* differences, hospitals can be compared relative to frontiers constructed both within-type (e.g. geographic region or function) and a combined frontier across type. Using this method average within- and between-type, efficiency can be estimated.

In general, whether *a priori* ordering of patient severity by hospital type is possible or not, methods of adjustment for predictive patient risk factors are required to eliminate economic incentives for cream-skimming of patients with lower risk for any given DRG in performance measurement. Methods of risk adjustment required to allow robust satisfaction of the common comparator assumption are considered in section 5.7 along with methods to allow robust satisfaction of coverage of effects in applying the correspondence theorem.

5.7 Satisfying correspondence theorem assumptions: a framework for appropriate incentives

The correspondence theorem explicitly makes the assumptions that:

1. disutility events cover effects of care in patients and;
2. hospitals face a common comparator, or differences in hospitals expected costs and disutility event rates (patient risk factors) can be adjusted for, to allow comparison, as if they faced the same comparator.

In allowing correspondence, an implicit assumption that the monetary value of effects of care can be considered as a constant willingness to pay, is also made, as it is in a net benefit framework (Stinnett and Mullahy, 1998).

In practice, to fully capture health effects attributable to care requires the ability to include negative and positive health effects from quality of care, within-admission, but also beyond-separation (to prevent incentives for cost and event-shifting). To satisfy the common comparator assumption requires compared hospitals to have access to the same technology, and differences in patient population risks of costs, and events prior to admission, are adjusted for. These conditions are, however, also requirements for appropriate peer grouping in avoiding incentives for cream-skimming. The net benefit correspondence theorem therefore provides a performance measurement framework for overcoming economic incentives for cost-shifting and cream-skimming and allowing economic incentives for quality of care consistent with maximising net benefit.

5.7.1 Satisfying the common comparator assumption

For a given clinical activity (DRG), the assumption that all hospitals compared face the same comparator can be violated if compared hospitals do not face the same:

1) expected risk of disutility events (e.g. mortality, morbidity) and costs in patients treated;
2) technology implicit in the frontier minimising costs and disutility events; or
3) prices of factor inputs.

5.7.1.1 Allowing for differences in patient risks within DRG

Patients presenting for the same DRG are not necessarily randomly distributed across compared hospitals, as they are across compared treatment arms with randomised control trial evidence used in health technology assessment.

An assumption of homogeneity in expected cost per patient, for each DRG across hospitals is implicitly made when comparing hospital performance with cost per case-mix adjusted separation. This assumption becomes an explicit assumption to be satisfied under the net benefit correspondence theorem when comparing performance with cost plus disutility events valued at WTP, as well as the assumption that hospitals face the same expected risk of disutility events.

To satisfy the correspondence theorem condition of a common comparator, various methods can be used to control for differences across hospitals in expected costs and effects at admission conditional on patient risk factors and other environmental differences that hospitals face. In adjusting for environment in any given clinical activity, the appropriate method will depend on the efficiency measure (technical, economic), method (Index, DEA, SFA) and data availability (predictive covariates and patient or hospital level cost and effect data)..

In the simplest case, considering performance measured with economic efficiency directly as costs plus disutility events valued as in net benefit, ordinary least square regression methods can be employed, with covariates representing predictive patient factors. Where individual patient data on cost and outcome of care (disutility event) are available, the linear nature of cost plus disutility events valued at WTP, and particularly the additive property of this statistic across patients, allows simple linear regression to be undertaken at a patient level, with this statistic as the dependent variable. Such an approach is analogous to the method illustrated by Hoch, Briggs and Willan (2002), adjusting for covariates and undertaking subgroup analysis in clinical trials by regressing on the linear net benefit statistic. Hoch, Briggs and Willan (2002) use the linear nature of net benefit to estimate a net monetary benefit (NMB) statistic for each individual (i) as:

$$NMB_i = l \times E_i - C_i, \qquad (5.11)$$

where:

? is the maximum acceptable willingness to pay per unit of health gain, and;

E_i and C_i represent the effect and cost of care for each individual i (i=1…n).

Using this statistic as the dependent variable for each individual, differences in net monetary benefit per patient for a treatment, allowing for patient covariates, were estimated by Hoch, Briggs and Willan (2002) with the regression:

$$NMB_i = a + \sum_{j=1}^{p} b_j x_{ij} + dt_i + e_i, \qquad (5.19)$$

where:

a is an intercept term;

t is a treatment dummy taking the value of 1 for treatment and 0 for control;

there are p covariates x, and;

e is a stochastic error term.

Net monetary benefit of treatment, relative to control and adjusted for covariates, can then be estimated by d.

Analogous to this approach in Hoch Briggs and Willan (2002), the linear form of costs plus disutility events, valued at threshold of avoiding disutility, would allow its use as a dependent variable in linear regression in comparing relative performance of hospitals adjusting for differences in patient covariates, where patient level data is available. That is, where individual patient data is available, regression of the same general form as equation 5.19 can be undertaken with (m-1) dummy variables for j=1..m hospitals compared. For each of these (m-1) dummy variables the regression can estimate an associated d_j value, relative to a reference hospital (value 0). In the case of minimising cost plus value of disutility events, the smallest d value (negative unless the reference hospital, in which case 0) would, however, represent the highest relative net benefit (best performing).

As in the case of Hoch, Briggs and Willan (2002), relative performance estimated using this regression method, would be conditional on the value of the threshold ceiling ratio via the effect of WTP in the dependent variable. Robust use of this method also requires patient level data on covariates, costs and effects. In Australia, patient level cost data is currently only routinely measured with micro costing methods in a sample of Victorian public hospitals with micro-costing, while routinely collected patient-level effects data linking to that beyond-separation was only available at the time of writing in Western Australia (Holman et al., 1999). However, neither of these data sets was accessible in undertaking this thesis.

Patient environment can also be adjusted for, with independent modelling of expected costs and risks based on patient factors under a frontier approach. If expected disutility

event rates, given observed patient risk factors, can be estimated from prognostic models, such as those from randomised control trials or routine data, various methods can be used to adjust for differences in baseline risk in a DEA approach. In using DEA with costs and disutility events as inputs, potential methods to allow for patient differences include:

1. Restricting comparison sets (peer grouping), to allow for *a priori* expected risk (as illustrated in 5.6.5);
2. Standardising for expected costs and effects, prior to performance measurement;
3. Allowing for expected costs and effects as non-discretionary variables in performance measurement and;
4. Regressing of efficiency scores on 'environmental' explanatory factors in second stage regression.

Peer grouping, allowing for *a priori* expected risk, demonstrated for hospital type in section 5.6.5, while improving on pure peer grouping, still reduces the comparison set, particularly for those hospitals with severe patient populations. The method also does not allow for difference in baseline risk within peer groupings, or between hospitals.

Expected disutility event rates can be estimated from logistic regression (Cox, 1972; Collett, 1994) using prognostic factors such as age, sex and co-morbidities at point of admission. A recent study by Simpson, Evan, Gibberd, Heuchean and Henderson-Smart (2003), provides an example of this at a clinical activity level, in comparing quality of care using routinely collected data. Intra-ventricular haemorrhage rates across hospitals, used as an indicator of quality of care in treating neonates in intensive care, were adjusted for routinely collected prognostic factors. Logistic regression was undertaken on the significant routinely collected predictive baseline factors of: sex; gestational age at birth; one minute Apgar score; antenatal corticosteroids and transfer at birth. This regression was then used to estimate expected risk of intra-ventricular haemorrhage, given observed covariates for each hospitals' patient population. A similar approach could also be employed to estimate expected costs for each hospital, based on covariates of predictive patient factors. To allow for interaction (covariance structure) between costs and effects,

regression of expected costs and disutility events should, however, be based on the same set of patient data.

Prognostic variables, representing expected cost and disutility event rates given covariates of patient populations, have a clear *a priori* direction of influence. A natural method, in input-orientated DEA, to minimise costs and disutility events, is to use these expected disutility events and costs as additional outputs (non-discretionary given an input- orientation). For example, the use of expected disutility events as an output ensures that comparison is relative to a theoretical hospital (linear combination of other hospitals) with equivalent, or greater, baseline risk. This approach provides an intuitive method of standardisation in comparison of relative performance amongst peers, given expected and observed disutility event rates.

A DEA study in hospitals by Morey, Fine, Loree, Retlaff-Roberts and Tsubakitani (1992), considered in greater detail in section 5.8, provides an example of adopting such an approach. Observed deaths were used as an input, and case-mix adjusted expected deaths as an output, in adjusting for baseline risk of mortality. A qualification to using this approach is that use of too many such non-discretionary variables can lead to overly restricted comparison. This was the case in the study of Morey Fine, Loree, Retlaff-Roberts and Tsubakitani (1992), where 12 such non-discretionary variables were used as inputs and outputs, resulting in only 168 of 300 hospitals in their study having any peers identified. However, where there are prognostic variables for expected disutility event rate and cost available, this approach could be employed without overly restricting comparator sets.

An alternative approach to adjustment for expected costs and disutility events, given patient risk factors, is to standardise cost per admission and disutility event rates prior to performance measurement. For example, standardising disutility event rates across hospitals might employ a three step process, such as:

(1) Calculating the ratio or difference of actual to expected disutility event rate;

(2) Applying the ratio, or adding the difference from step (1) in each hospital to the mean rate across all hospitals to estimate an adjusted rate, without differences due to baseline risk and;

(3) Applying the rate from (2) to admissions for each hospital, to estimate standardised disutility events.

The ratio (relative risk) alternative ensures that event rates cannot be negative, a potential problem with using differences to standardise in performance measurement. While such prior standardisation imposes no restriction on comparators, the trade-off is that it does not use the implicit observed structure of technology in adjustment, and can make interpretation of performance scores difficult. In comparison, including expected costs and disutility directly as outputs in performance measurement has an intuitive and transparent interpretation and implicitly allows for technology in adjustment, to the extent that it is represented by observed practice.

Second stage regression of DEA derived technical efficiency scores (which lie between 0 and 1) on environmental variables can also be considered as a method for identifying and adjusting for significant environmental factors not considered in efficiency measurement. In undertaking such second stage regression on efficiency scores, Tobit regression undertaken with maximum likelihood methods is preferred over ordinary least squares methods, in explicitly allowing for the censored nature of technical efficiency scores at 1. As described in appendix 5.1, this method allows for the effect of covariates in relation to likelihood of being on the frontier, as well as degree of inefficiency. Technical efficiency scores are simply transformed into technical inefficiency scores using (1/efficiency-1). This represents a convenient normalisation for estimation purposes, for which, under constant returns to scale, the transformed inefficiency score can be interpreted as the percentage increase in outputs possible with best practice.

Puig-Junoy (1998) provide an example of a second stage regression approach in a DEA study of neonatal intensive care performance at a clinical activity level, allowing for

expected survival in second stage regression. Puig-Junoy (1998:267) in reference to the need to adjust for expected survival, argued that:

> "A major problem in efficiency analysis of health care providers is the difficulty of measuring the patient in the input (severity of illness) and output set (improved health status)".

Days of survival and discharge status, were used as outputs in estimating relative technical efficiency of intensive care units, which were then regressed in second stage on exogenous explanatory variables, including probability of survival at admission.

While such second stage regression approaches are technically feasible, it should, however, be noted that they can be seriously biased where there is a relationship between variables in first stage DEA, and variables in second-stage regression. In the case of baseline and actual patient risk of survival in the study of Puig-Junoy (1998), the assumption of independence between these variable is likely to be seriously violated, and would be with any such variable representing expected cost or disutility event rate. In comparison, an approach using expected survival directly as a non discretionary input, avoids such biases, while appropriately, but not prohibitively, restricting comparators (as in the analysis of Morey et al (1992), where 12 predictive non-discretionary variables were used).

In summary, in adjusting for differences in patient risk, where patient data is available in compared hospitals, direct adjustment for covariates on costs plus disutility events valued as in net benefit is preferable in considering economic efficiency. Where patient level data on costs and effects of care of compared hospitals is unavailable, but expected costs and disutility events based on patient covariates can be estimated from prognostic models[5] (based on other hospitals cost and effect data), direct inclusion as outputs in DEA is suggested to offer the best alternative. This approach offers an intuitive and transparent method of adjustment for variables with clear direction of influence. It can

[5] Models of expected costs and effects from patient level data should be based on the same hospital's data to allow estimates implicitly based on their joint distribution.

adjust for differences in patient mix (risk factor profile), whether endogenous determined (and cream skimming incentives are important to prevent) or exogenous to the hospital (and hospitals should not be punished for factors outside of their control).. It also makes use of implicit technology from observed practice, and should not overly restrict comparator sets if limited to non-discretionary variables of expected costs and disutility events. In adjusting for patient differences in funding, prior standardization based on expected relative to observed effects is required, as described in section 6.7.

In performance measurement, adjustment for differences in expected costs and effects of patient populations is important to deter incentives for potential selection bias (cream-skimming) in choice of patients with lower expected costs and risk of disutility bearing events. However, such adjustment is also important in not rewarding or punishing hospitals for factors outside of their control. Environmental factors such as geography, regulation, or differences in technology or services (teaching, research) may be .outside of the control of the provider.

Peer grouping, second stage regression and other approaches, outlined in Coelli, Rao and Batesse (1998:166-171), can be used to explore and adjust for effects on efficiency of such purely environmental factors. In adjusting for such environmental variables, Coelli Rao and Batesse (1998:171) note that a two stage approach, with second stage regression, will often be the best of these alternatives. It has advantages of a lack of *a priori* assumptions, the ability to include categorical and continuous variables, and to undertake tests of the significance of effects.

5.7.1.2 Allowing for technology differences

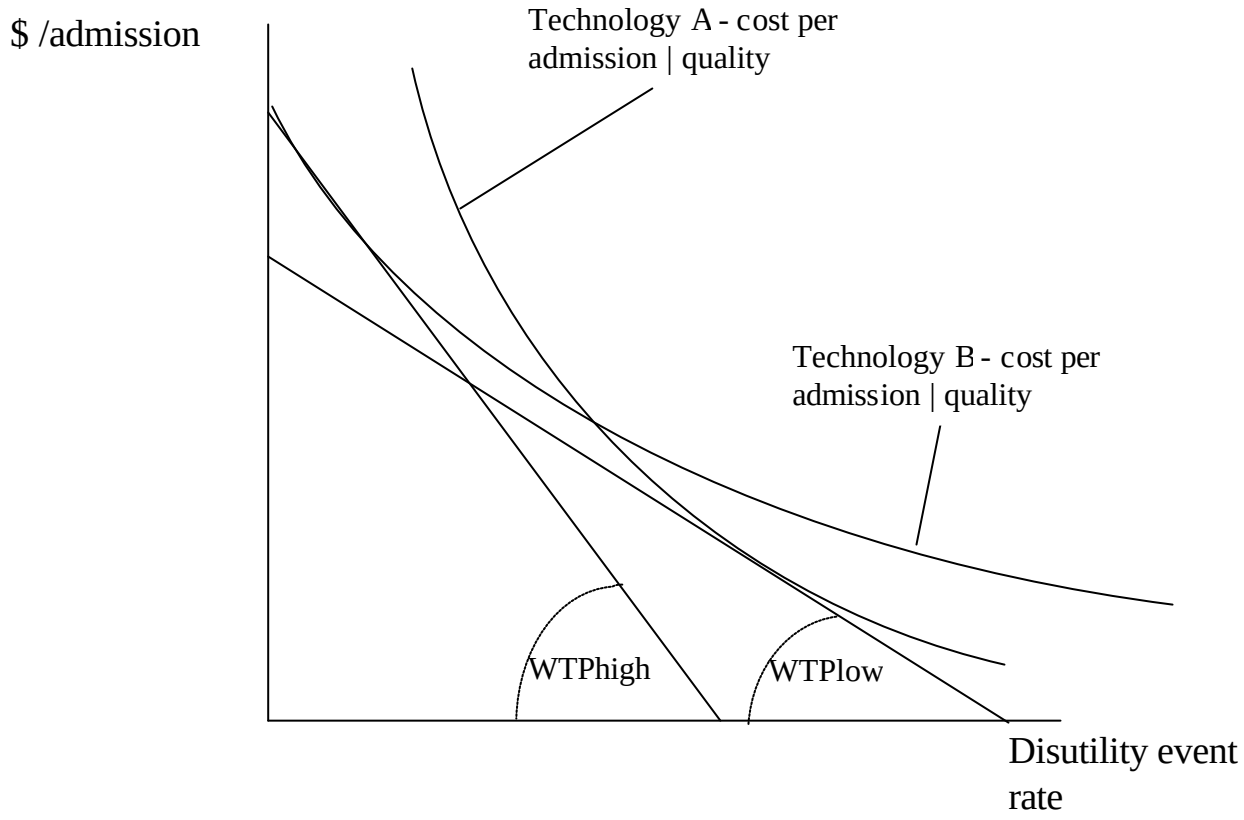
If access to technology differs between hospital types, peer grouping can be undertaken in determining comparator sets when assessing relative performance of hospitals with different technology. If common quality of care indicators' are used across hospitals with different access to technology, then analogous to health technology assessment, use of technologies in practice can be compared in considering dominance or preferences at a

given WTP. In making such comparison, adjustment for differences in risk for patient populations is, however, likely to be important, as outlined in section 5.7.1.1.

For example, depending upon DRG, greater access to new medical technology in principal referral hospitals, may reflect greater complexity of patients, and hence baseline risk of disutility events and expected costs. Use of different technologies by hospitals for the same DRG may, therefore, appropriately reflect differences in severity of patient populations.

Where differences in expected cost and disutility event rates are adjusted for, relative performance of technologies, as well as individual hospitals, can be assessed. Comparison of best practice in use of technologies can be assessed in comparing the relative position of isocost curves (at the decision maker's value for avoiding disutility events) tangent to best practice frontiers for each technology. For example, in figure 5.8, best practice with technology A is preferred to that of technology B with low willingness to pay (to avoid disutility events), but best practice for B is preferred to A with high willingness to pay.

Figure 5.8: relative performance of hospitals with different technologies conditional on willingness to pay



Comparison of average performance across technologies in practice may be seen as more robust than identification of best practice hospitals at an individual DRG level hospitals (where comparison can be significantly affected by sampling variation[6], as described in chapter 9). This may particularly be the case for hospitals with a small number of admissions for a clinical activity, and hence a greater intrinsic uncertainty and likelihood of being on the frontier with low disutility event rates by chance.

Aggregating across all hospitals, rather than best practice hospitals in isolation, allows more robust consideration of relative performance of technologies in practice. Relative performance of technologies, in practice, can then be considered in comparing average cost per patient and disutility event rate, aggregated across hospitals by their use of alternative technologies. The relative position of isocost curves (with slope representing the value of avoiding disutility events) passing through these points allows relative performance to be ordered consistent with net benefit maximisation. Where WTP is unknown, performance could be conditioned on potential values of WTP, and regions over which technologies may be preferred can be identified.

Incremental consideration of frontiers in the cost disutility plane can also be used directly in health technology assessment, where there are multiple strategies or technologies. The use of frontiers to discriminate between technologies in cost disutility space is described and illustrated in chapter 8 The assumption of a common comparator is shown to be naturally satisfied by use of randomised control trial evidence in assessment of treatment options. Reframing measured effects of incremental survival rate, reduction in morbidity rate or life years and QALYs saved as disutility events allows the assumption of coverage of relative effects of care to be satisfied. Incremental rates of survival and reduction in morbidity become incremental rates of mortality or morbidity. Life years or QALYs

[6] Sampling variation may also be problematic in interpreting relative performance, and particularly for hospitals with smaller number of admissions; aggregation across hospital activities reduces this role of chance. Reducing the role of chance aggregation with industry cost shares by activity, as outlined in chapter 3, would enable performance at a hospital level to be assessed, without paradoxical comparison problems. Bayesian shrinkage estimation and stochastic frontier analysis also offer the potential to allow for such sampling variability, as described in future directions in chapter 9.

saved per patient translate to life years or QALYs lost relative to the most effective strategy.

5.7.2 Satisfying disutility events capturing effects of care

In measuring the relative effects of quality of care across hospitals, not all disutility bearing effects will be observed within hospitalisation. This is particularly important, given discretion of providers with respect to point of separation, and hence scope to act on incentives for cost and effect shifting, present if such practices are measured as 'improving' performance. These perverse incentives can be overcome, either by including post-separation effects attributable to care, or by attempting to measure effects of care with health status at point of separation. In either case, to be comprehensive in coverage of effects[7] ideally requires the ability to include utility-bearing effects of care, reframed as disutility events.

5.7.2.1 Including utility bearing aspects of care as disutility events

In considering whether health effects can be represented by disutility events, the World Health Organisation's definition of health (WHO, 1948) is important to consider:

> "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity."

An inability to measure, or proxy, for utility bearing aspects of quality would only allow a partial perspective of health effects of care.

The ability to include utility bearing aspects, related to physical, mental and social functioning, is likely to be particularly important in sub-acute and non-acute hospital activities, where disutility events may be inferior indicators of quality of care.

Not meeting a standard can be used as a disutility bearing event, where utility-bearing aspects of care can be measured as a standard, such as a required level of physical, mental or social functioning. The disutility event rate then represents the proportion of the patient population not meeting a standard. More generally, in considering relative

[7] and apply the correspondence theorem consistent with net benefit maximisation

performance, as the correspondence theorem demonstrates, minimising the negative value of disutility event rates provides the same ordering[8] as maximising the value of equivalent positive effects of care. Reframed as disutility event rates, utility bearing aspects of care can be used in measuring relative performance of hospitals applying the correspondence theorem. To allow for differences in patient prognostic factors, in satisfying the common comparator assumption, adjustment of rates could be undertaken either by adjusting:

(1) the standard of functioning itself for observable patient risk factors, in defining whether it is met or not, for a given patient or;
(2) the rate meeting a fixed standard across all patients.

The first approach may be preferred, given that it can allow for patients capacity to benefit in meeting a standard (a concept explored further in section 8.2.2).

While not meeting a defined standard allows inclusion of utility bearing aspects of outcomes as disutility events, the incentives created by such standards are only present up to the threshold level the standard identifies. To allow a finer distinction, where cardinal measures of health status with defined endpoints or health related utility can be estimated, equivalent disutility event rates are average disutility or average limitation in functioning.

For example, if average functional ability after surgery was 50% and 60% in two hospitals, functional limitation would be 50% and 40% respectively. Similarly if average health related utility was 0.6 and 0.8 in two hospitals, equivalent average disutility would be 0.4 and 0.2. Utility bearing levels of functioning and health can, therefore, be naturally reframed as disutility events in considering relative effects across providers if they are measurable as either: meeting standards, health related functioning, or health related utility.

[8] In chapter 6 the correspondence is also shown to hold for differences between maximising net benefit and minimising cost plus disutility vents valued as in net benefit.

To avoid the potential for finessing by providers, the level of disutility, functional limitation or attaining of standards need to be verifiable, just as it would be required to be in a utility bearing framework. At a clinical level, clinical audit and peer review of a random sample of admissions could be undertaken to monitor and create disincentives for finessing, explored in detail in policy implications in section 7.4.2.

5.7.2.2 Including process aspects of care

The relative value of technologies in a net benefit framework is measured by health effects. However, it should be noted that in health care, valuing output of care as health effects of care alone (for example as argued by Culyer, 1992) is not without challenge. As Mooney (1994a, 1994b) points out, this is a narrow, consequentialist view of the value of care in which it is assumed that the only output is health. Performance assessment would ideally extend beyond value of health outcomes, to include relative value from process of treatment.

In applying the correspondence theorem to performance measurement, process as well as outcomes could be included in measuring quality and relative performance of care, where they can be translated into not meeting standards, disutility bearing aspects of care measured cardinally, or disutility directly. Application of the correspondence theorem can therefore move beyond health effects and allow the model to include quality of services in the dual sense of health outcomes and process, or functional characteristics of quality of service, proposed by Gronroos (1984). Under Gronroos' model, quality is divided into technical quality related to outcomes, and functional quality reflecting the way services are delivered. Functional quality represents (dis)utility bearing aspects of the interaction between the customer and the organisation, such as waiting time or time of staff spent with patient. Functional quality under this definition recognises that service is a process of interaction between customer and organisation (patient and health service provider), in which the technical dimension does not fully represent the overall quality of the service as experienced by the customer.

5.7.2.3. Health related utility and average cost effectiveness versus disutility and net benefit maximisation

As a result of reframing health related utility bearing aspects of care as disutility events and specifying them as inputs, the implicit objective function in performance measurement can correspond with net benefit maximisation, under correspondence conditions. In comparison, if health effects are specified as utility bearing outputs (for example survival or average health related utility at separation), and compared relative to costs of care, the implicit objective function in performance measurement is minimising average cost effectiveness.

In considering relative performance of hospital services across populations, average cost effectiveness in health care is problematic as a relative performance measure, given the incremental and no-tradable nature of health effects of care. As described in section 4.2, the effects of care on health are incremental, given floor effects in health outcomes of care across patient populations. Under these conditions, the need to compare with an appropriate comparator is well established and is the basis for incremental consideration of costs and effects in the cost effectiveness plane in health technology assessment (Drummond, O'Brien, Stoddard and Torrance, 1997). Similarly, in comparing hospital performance, even if hospitals did nothing, the health effect would not be 0 (which assumes all patients die). A lowest cost effectiveness ratio can represent lowest net benefit (worst performance) and even net loss in health effect, relative to doing nothing.

Even if effects could be measured for each hospital as an incremental change in health-related utility, standardised for baseline risk, the lack of tradability of health effects implies a lower cost per incremental effect will not necessarily be preferred. Where there is a higher absolute health benefit of care, a higher cost per health gain may be preferred. The non-tradeable nature of health implies that health effects are not able to be factored up in the same way as other goods. Once a patient population has been treated, it is not the case that more health can simply be bought for that patient population. In underlining this, it is worth repeating a quote of McGuire, Henderson and Mooney (1988) from section 4.2:

> "Health itself is not tradeable in the sense it cannot, strictly, be bought or sold in a market... health is not exchangeable." McGuire, Henderson and Mooney (1988:32)

In measuring relative performance, what is meaningful in ordering relative performance is trading-off incremental health gain against incremental cost relative to a threshold value. This is implicit in performance measured consistent with maximising net benefit per admission. As the correspondence theorem demonstrates, specifying disutility events as an input under correspondence conditions explicitly allows an incremental trade-off between costs and effects in relative performance measurement, consistent with maximising net benefit. However, specifying disutility events as an output does not.

If utility gain can be measured, then specifying 1 less average health utility gain as an input at the decision makers WTP allows relative performance to be more meaningfully represented than specifying average utility gain as an output. Under correspondence conditions, this allows performance measurement consistent with net benefit maximisation, whereas, an output specification, at best, represents an underlying objective function of minimising average cost per health gain.[9]

In the absence of incremental health gain measured as a non-negative variable, the advantages of an input specification over an output specification are starker. Performance, measured with these alternative input and output specifications reflect an underlying objective function of maximising net benefit per admission, versus average cost effectiveness, respectively. The following theorem illustrates how performance, measured with an output specification reflecting average cost effectiveness differs from net benefit maximisation (consistent with an input specification of health disutility under correspondence conditions).

[9] The measurement of health gain implies that the common comparator assumption has been satisfied in adjusting for baseline effects of care.

5.7.2.4: Lack of correspondence between specifying health outcomes as outputs and maximising net benefit

Lack of correspondence theorem: A correspondence does not exist between ordering hospital performance, maximising net benefit per admission and minimising cost per unit of health effect, valued as in net benefit.

Proof by contradiction

Assume there was a correspondence between maximising net benefit and minimising average cost per unit of health effect. Then the following relationship is required to hold in any bilateral comparison with a common comparator:

$$k \times E_i - C_i > k \times E_j - C_j \Leftrightarrow C_i/(k \times E_i) < C_j/(k \times E_j) \quad (5.13)$$

Now consider the LHS of this equation (5.13) where i does not dominate j, and hence incremental effect and incremental cost of i relative to j have the same sign (that is there is a trade-off between incremental costs and effects of i and j).

$$k \times E_i - C_i > k \times E_j - C_j$$

$$\Leftrightarrow k \times (E_i - E_j) > C_i - C_j$$

$$\Leftrightarrow \quad k > \frac{(C_i - C_j)}{(E_i - E_j)}, E_i - E_j > 0 \text{ , or;}$$

$$k < \frac{(C_i - C_j)}{(E_i - E_j)}, E_i - E_j < 0.$$

Ordering of net benefit depends on k.

Now consider the RHS of the assumed correspondence. As k is a common factor and k>0, the ordering of average effect per cost is independent of k.

$$C_i/(k \times E_i) < C_j/(k \times E_j) \Leftrightarrow C_i/E_i < C_j/E_j$$

Therefore, the correspondence does not hold for any bilateral comparison where $C_i/E_i < C_j/E_j$ and either:

$$k < \frac{(C_i - C_j)}{(E_i - E_j)}, E_i - E_j > 0 \text{ ; or}$$

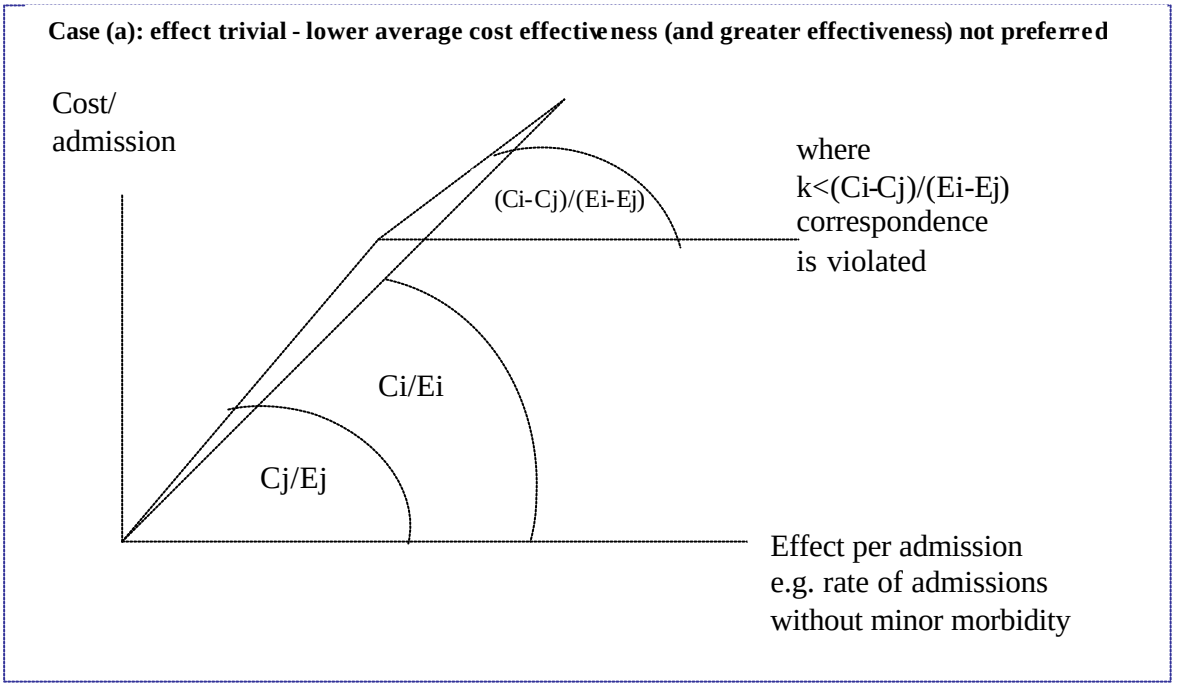
$$k > \frac{(C_i - C_j)}{(E_i - E_j)}, E_i - E_j < 0.$$

Illustrated in the incremental cost-effectiveness plane, the correspondence is violated where average cost per effect is less for hospital i than j and either:

(a) the average effect in hospital i is greater than that in j but the incremental cost-effectiveness ratio for i relative to j is greater than the threshold value (figure 5.9(a)) or;

(b) the average effect of i is less than that of j but the incremental cost-effectiveness ratio for j relative to i is below the threshold value (figure 5.9(b)).

Figure 5.9: lack of correspondence - net benefit and average cost effectiveness



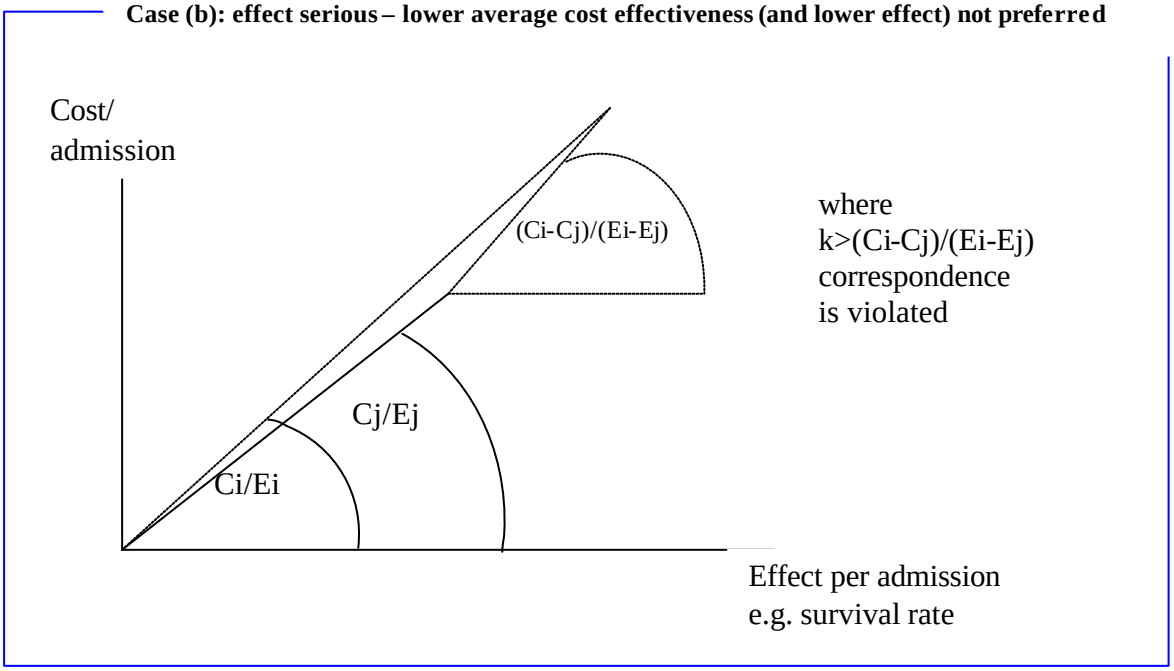


Figure 5.9 illustrates the familiar finding from health technology assessment that comparing the relative values of average cost per unit effect does not necessarily correspond with preferences based on accepting incremental cost-effectiveness ratio below a threshold (Weinstein and Fineberg, 1980; Drummond, Stoddard and Torrance, 1989), or equivalently, maximising net benefit.

Using average cost-effectiveness, in case (a) depicted in figure 5.9(a), relative performance is overstated for hospitals with low disutility event rates relative to those with high disutility event rates. Characteristically, this discrepancy can arise if the WTP to avoid disutility events is low (effects of care minor) compared with costs of care. Conversely, in case (b) depicted in figure 5.9(b), relative performance is understated for hospitals with low disutility event rates, relative to those with high disutility event rates. This is the more likely case in hospitals, and characteristically arises if measured effects of care are serious and WTP to avoid disutility events are large, relative to costs of care.

In general, ordering of performance with average cost-effectiveness does not correspond with incremental cost-effectiveness below a threshold. Even if all effects of care could be measured as utility bearing aspects, performance measurement would still reflect a more appropriate objective function if effects were reframed as disutility bearing events and included as inputs.

5.7.2.5 Allowing for post-hospitalisation effects

Appropriate linked data across hospitals can be simply incorporated within the proposed correspondence framework, by including additional cost and disutility events beyond-separation. In determining the appropriateness of linkage, related questions which need to be addressed for any given DRG include:

1. What (proportion of) effects are attributable to care received in admission, as opposed to external environment beyond-separation?
2. At what proximity to point of separation are costs and events attributable?

At a clinical activity level, decision analytic modelling methods used in health care (Weinstein and Fineberg, 1980; Pettiti, 1994; Hyunink et al., 2001) can be used to inform the process of identifying beyond-separation health effects and costs of associated care. Ideally, if data from linked events allows point of incidence post-separation to be estimated, weights can be applied that dampen with time, post-separation. Post-separation disutility events and costs can then be attributed to reflect effects of within-hospitalisation quality, as opposed to external factors beyond-separation. Weights could also be adjusted for baseline patient characteristics or other exogenous explanatory factors. For example, for death registry data, cause of death could be used, while for readmission or care in other settings, cause codes or case notes could be considered.

5.7.2.6 Valuing disutility events

To enable correspondence with maximising net benefit (Stinnett and Mullahy, 1998), values placed by decision makers' on combinations of disutility events need to correspond with those for effects of care. For decision makers to systematically value disutility events, the equivalent incremental effect of avoiding disutility events would ideally be estimated in a common metric, such as quality adjusted life years.

In estimating expected incremental effects and costs attributable to disutility events, decision analytic methods can be employed, as in health technology assessment. Evidence from clinical trials, epidemiological data (such as life tables for patient populations) and costs of current practice can be used to populate decision trees, modelling expected effects and cost of further care with different combinations of disutility events.

For clinical activities where evidence to populate decision trees is not currently available, decision analytic methods, in combination with the correspondence theorem, create a framework to identify what is required for quality of care to be appropriately included. This may include appropriate data linkage, as well as identifying effects within-care and identification of risk factors.

In clinical activities where there are multiple types of disutility events, decision analytic methods can be employed to identify incremental effects on QALYs from different combinations of effects. These values can then be applied to patient populations separated by combinations of events. This method allows appropriate interaction of disutility events and their effects on costs and outcomes in valuing within a decision analytic framework. Such an approach is preferable to making assumptions, such as: the valuation of avoiding disutility events is additive.

A decision analytic approach, separating populations by combinations of events in valuation, avoids double counting of costs or benefits in specifying alternatives and their consequences. Such a systematic approach addresses concerns raised by Drummond, O'Brien, Stoddard and Torrance (1997), Diener, O'Brien and Buxton (1998) and Currie Donaldson, O'Brien, Stoddart, Torrance and Drummond (2002) over the disembodied, and consequent variable, use of contingent valuation methods in health care programs. While their concerns related to use of such methods in health technology assessment in estimating money values for non-marketed program benefits, the same issues arise in valuing relative non-marketed effects between providers of public hospital services.

As in health technology assessment, use of a net benefit framework (Stinnett and Mullahy, 1998) in performance measurement, implicit in applying the correspondence theorem, challenges policy makers to identify monetary values for relative effects of care. Policy makers setting values for effects of care in activities under their control is important. In the absence of these values, and transaction conditions for patients preferences to be reflected in provider behavior, the vacuum is filled by variable economic incentives for quality of care that fail to appropriately trade off the cost and effects of care in practice. Economic incentives for net benefit maximisation are not created without explicit values for effects of care. Where an economic value for effects of care is defaulted on altogether, as in the case with case-mix funding, the lack of accountability provides scope for technical, as well as allocative, inefficiency.

Including value of disutility bearing events as inputs, can be seen conceptually as analogous to inclusion of costs of payouts for negligence where disutility events were compensated for. However, an approach directly including such compensation payouts would not represent a societal perspective, and has numerous potential biases. Biasing factors include:

1. imperfect knowledge of quality of care and access to legal representation for patients or their next of kin;
2. differences in insurance coverage of hospitals;
3. the need to include the legal costs of defence as well as payouts across insured and non-insured;
4. delay or exclusion in obtaining settlement data;
5. whether payouts reflect decision makers' value of quality and;
6. differences in payouts in jurisdiction, given differences in law or its interpretation.

In comparison, applying a decision maker's threshold value of willingness to pay to reduce disutility bearing event rates is appropriate in trading off effects, and costs, of quality of care across patient populations.

While decision makers' need to identify relative monetary values of effects of care (avoiding disutility events) in applying the correspondence theorem, relative values across health systems, clinical activities and different effects of quality of care should reflect differences in ability to pay and preferences. In reflecting decision makers' values the correspondence theorem can be applied in comparison of relative performance within, but not necessarily across, health systems or jurisdiction. In general, in applying the correspondence theorem to relative performance measurement, common relative decision making values for effects of care are required. Once these relative values are established within any health system, the absolute value of avoiding disutility events would ideally be determined universally under constrained optimization, reflecting the ability to pay for quality of care. This gives the flexibility to reflect values which can differ in relative terms across cultures, as well as in absolute terms, based on ability to pay.

5.7.2.7 Valuing process aspects of care

In valuing process aspects of care, discrete choice conjoint analysis can allow willingness to pay (WTP) for process attributes, as well as health outcomes, to be elicited where cost is included as an attribute (Ryan, McIntosh and Shakley,1998a). Using responses to discrete choices between alternative scenarios, relative importance of attributes and their marginal rate of substitution can be estimated. From these marginal rates of substitution, willingness to pay can be directly estimated where trade-offs for monetary aspects are included.

Conjoint analysis was recommended to the UK treasury in evaluating quality in public service provision generally by Cave, Burningham, Buxton, Hanney, Pollitt, Scanlan and Schurmer (1993) following successful use in environmental (Opaluch, Swallow, Weaver, Wessels and Wichens, 1993) and transport settings (Wardman, 1988).

In application to health care, conjoint methods have been used in eliciting patient values and preferences, including those for:

1. waiting times (Propper 1990, 1995) and waiting, travel times and complication rates (Jan, Mooney, Ryan, Bruggerman and Alexander, 2000);
2. Management aspects including continuity of contact with staff and health outcomes for miscarriage (Ryan and Hughes, 1997) and invitro-fertilisation (Ryan, 1999);
3. Magnetic resonance imaging in knee injuries (Bryan, Buxton, Sheldon and Grant, 1998);
4. Blood transfusion support (Van der Pol and Cairns, 1998);
5. Attributes of primary care (Ryan, McIntosh and Shakley, 1998b) and;
6. The doctor-patient relationship (Vick and Scott, 1998).

To the extent that decision makers' willingness to pay threshold values reflect patient preferences, such values could be used in informing relative values in including process aspects of care under the correspondence theorem. In estimating robust relative values

using conjoint analysis, further research on modelling interaction between values of process and outcomes of aspects of care is, however, required.

To adequately allow for interaction or conditioning of valuation for different combinations of attributes, more sophisticated nested models such as that used by Goldberg (1995), rather than linear additive models currently used in health care applications, are required. While such refinements are likely to be critical to the robustness of valuations, the increase in data requirements of such nested approaches can be considerable. As Boyle and Bishop (1988) noted in comparing the relative merits of contingent valuation eliciting methods generally, take-it-or-leave-it discrete-choice methods increase sample size requirements to allow statistical validity. The feasibility of undertaking conjoint analysis with nested models may, therefore, be problematic in many circumstances due to cost considerations. This would particularly be the case using discrete choice methods with interview, as recommended by the National Oceanic and Atmospheric Administration (NOAA, 1993). In choice of valuation elicitation method discrete choice take-it-or-leave-it approaches were, however, recommended by the NOAA, based on advantages of avoiding starting point framing biases of alternative approaches, such as bidding games.

Framing effects of methods such as bidding games could, however, be avoided in eliciting preferences using random starting points and computer-based elicitation methods, as Diener, O'Brien and Buxton (1998:323) suggested. Even without random starting points, O'Brien, Goeree, Gafni, Torrance, Pauly, Erder, Rusthoven, Weeks, Cahill and LaMont (1998) empirically found no starting bias in comparing bidding algorithms in a study of the feasibility of contingent valuation in managed care.

A general question raised by O'Brien and Gafni (1996) and Diener, O'Brien and Buxton (1998) relates to framing effects in valuing willingness to pay (for increase in consumer surplus or utility) versus willingness to accept (decrease in consumer surplus or utility). The NOAA, in its guidelines (NOAA, 1993) on use of contingent valuation in environmental damage assessment for cost benefit analysis, recommended WTP over

WTA due to its more conservative estimation of elicited value. In valuing disutility events when comparing relative performance of hospitals, use of willingness to pay, while allowing correspondence without knowledge of comparator effects, gives the benefit of doubt to hospitals with low quality care, as is described in detail in section 5.7.3.

However, framing questions to estimate willingness to pay in contingent valuation can be problematic in health care settings, as Diener, O'Brien and Buxtin (1998:323) noted:

> "losses from consumers and some form of WTA may be unavoidable.''

5.7.2.8 A summary of covering effects with disutility events

While the ability to satisfy coverage of the relative value of health effects by relative value of disutility events in the correspondence theorem is constrained at aggregate level, it does not appear technically limited at a clinical activity (DRG) level. Using decision analytic methods, relevant effects of care can be flexibly and comprehensively identified at a clinical activity level, including those beyond treatment. Health effects can be directly included or reframed as disutility events whether they are:

1. rates of morbidity;
2. rates of mortality;
3. rates of not meeting a standard of health-related functioning;
4. average limitation in health-related functioning measured on a cardinal scale at point of separation or;
5. average health-related disutility at separation.

In valuing combinations of disutility events, decision analytic methods can be used to identify equivalent net effects in a common metric such as QALYs, or using methods such as discrete-choice conjoint valuation, where disutility event rates represent aspects of quality other than health effects. In measuring relative performance of providers, process aspects of care, such as waiting times and standards of care, can be framed as disutility events, where appropriate, and be valued alongside health effects, informed by methods such as conjoint analysis.

5.7.3 Willingness to pay versus willingness to accept

In allowing the choice of comparator to be arbitrary in correspondence, an implicit assumption is required that the value of avoiding disutility events is constant between hospitals at different levels of disutility. Characterized as the willingness to pay in a net benefit framework (Stinnett and Mullahy, 1998), the value of avoiding disutility events is expected to be greater than WTP at disutility rates above that of next best alternative care, with loss aversion under prospect theory (Kahnemann and Tversky, 1978), as described in section 4.6.2. The value of willingness to accept perceived health losses is estimated to be 2-3 times greater than willingness to pay (WTP) for the same health gain (Willan, O'Brien and Leyva , 2001).

Use of willingness to pay, therefore, gives the benefit of doubt to providers of low quality (high disutility event rate) care. If provider quality of care results in expected effects of care below that of a common comparator, they have net benefit overestimated (net loss underestimated), and economic efficiency overestimated. This can only be adjusted for if the value of WTA and a comparator disutility rate is known.

The effect of being able to adjust would be to further reinforce divergence of relative performance ordering, using WTP in the proposed approach, from that where quality is valued at 0. In practice, use of WTP rather than WTA cannot, therefore, be seen as a limitation of the proposed method relative to comparison with current performance measures and incentives. The need to move incentives from cost minimisation to net benefit maximisation is further reinforced by the risk of health losses.

5.8 Previous modelling of disutility events as inputs

Modelling disutility bearing events as a conventional input, rather than an undesirable output, in DEA in measuring technical efficiency is not strictly new. Pittman (1981), Cropper and Coates (1992), Haynes et al (1993, 1994) and Rheinhardt, Lovell and Thijjsen (1999) have all considered efficiency including undesirable by-products such as pollutants and waste as inputs. As Pittman (1981) stated and Rheinhardt, Lovell and

Thijjsen (1999) re-iterated: "the relationship between an environmentally determined variable and output looks like the relationship between conventional input and output."

Despite this, in hospital performance measurement, the only example I am aware of where disutility events have been modelled as inputs is a study by Morey, Fine, Loree, Retlaff-Roberts and Tsubakitani (1992) of 300 US hospitals. This study considered trade-offs between cost and quality of care at a hospital level using a variable returns to scale DEA formulation, where quality was represented by actual, relative to expected, deaths, given DRG case-mix. Deaths were specified alternatively as either:

1. a non-discretionary input in minimising expected cost or;
2. a discretionary input, in minimising mortality rate, given costs.

Comparator sets for hospitals were, however, severely constrained, in specifying a variable returns to scale formulation and including non discretionary outputs including:

> case-mix weighted discharges; normal births; abnormal births; rehabilitation days care; sub acute patient days; visits to emergency rooms; clinic visits; ambulatory surgeries; and medical education expenditure.

Target hospitals were restricted to hospitals that produced at least the same level of all these types of outputs with the same, or no more, expected mortalities or beds available. Under these conditions, only 168 of the 300 hospitals in the study were reported to have an appropriate peer group identified.

This demonstrates the problem, at a hospital level, of requiring specification of too many outputs to avoid aggregation of heterogeneous clinical activities. This heterogeneity problem was stressed by Newhouse (1994). In critically assessing the use of frontier methods in performance at a hospital level, Newhouse (1994:321) portrays the bind of either not having enough sample to enable discrimination of performance in more than 500 diagnostic related group activities (even with 5000 hospitals in the USA), or measurement error and lack of interpretability in aggregating outputs.

A clinical activity (DRG) level of analysis, as noted in chapter 3, can overcome the problem of heterogeneity in outputs at a hospital level, while also flexibly allowing quality indicators appropriate to clinical activity (for example specific types of iatrogenic event, readmission, functional limitation or morbidity) to be considered. As Morey et al. (1992:678) admitted in noting limitations of their analysis (bracketed text added):

> "Other measures (than hospital deaths) such as risk-adjusted rates of re-admission, risk-adjusted post surgical complications and risk-adjusted mortality rates within specific periods following admission are needed."

While Morey et al. (1992) recognised a trade-off between costs and reduction in death rate (as an indicator of quality of care), a theoretical basis for including disutility events as an input was not identified or explored. The appropriateness of an objective function specifying disutility events as an input was not identified. In fact, nowhere in the text was it indicated that deaths were modelled as an input. This could only be implied from the variable returns to scale DEA cost-minimising formulation presented in appendices in their publication (Morey et al, 1992:697).

The correspondence theorem identified in this thesis provides the theoretical support for specifying disutility events as inputs to representing effects or quality of care consistent with net benefit maximisation. It represents the first illustration of a correspondence between an input specification of disutility events and net benefit maximisation.

At an individual activity level of analysis, use of decision analytic methods allows robust and comprehensive inclusion of disutility events as quality of care indicators in satisfying correspondence assumptions. In peer identification non-paradoxical aggregation from a clinical activity level is also possible, as chapter 3 addressed in detail.

5.9. Summary

A framework and method for measuring relative performance of hospitals at a clinical activity (DRG) level, including quality of care consistent with maximising net benefit per admission has been developed and illustrated in chapter 5. This framework is based on a

one-to-one correspondence between maximising net benefit and minimising costs and disutility events per admission (valued at decision maker's threshold), assuming:

(1) effects of care are captured by disutility bearing events and;

(2) hospitals face a common comparator (costs and effects of compared hospitals can be adjusted for differences across hospitals in patient risk factors).

The correspondence has been demonstrated to allow relative performance measurement consistent with maximising net benefit, using currently collected cost and disutility event data (such as mortality, morbidity and readmission). Applying the correspondence theorem to performance measurement using frontier method of DEA has been illustrated with Australian data at a clinical activity (DRG) level to allow policy makers to be informed of:

1. Relative performance conditional on value of avoiding disutility events (quality of care) consistent with order in maximising net benefit per admission;

2. Technically efficient hospitals as those that minimise cost conditional on effects of care (disutility events per admission), independent of the monetary value for effects of care;

3. Regions of value for effects of care (WTP) over which technically efficient hospitals maximise net benefit, using a back solving method;

4. Decomposition of economic efficiency corresponding with net benefit maximisation into technical efficiency minimising costs and disutility events, and allocative efficiency in valuing effects of care, and, where appropriate, scale and congestion efficiency;

5. Industry shadow price of avoiding disutility events at WTP where cost-share weighted economic efficiency is maximised.

Illustrations in this thesis have been noted as limited by data constraints faced in adjusting for:

1. differences in expected costs and mortality rates of hospital, requiring the assumption of a common comparator and;
2. differences in effects of care beyond-separation.

However, in general methods have been identified for adjust for differences in risk factors at admission, and effects post admission, and the data and level of analysis required to robustly apply these methods. Data linkage to effects beyond separation and patient level data on costs and effects for risk adjustment are required to robustly adjust and satisfy correspondence assumptions. At a DRG level, disutility rates can be adjusted for without confounding across activities at aggregate levels, and quality indicators are flexibly and appropriately identified, using decision analytic methods. In general the linear framework provided by the correspondence theorem framework allows simple and intuitive adjustment with the ability to include multiple disutility bearing events in covering effects of care, including those attributable beyond-separation.

In satisfying coverage conditions of the net benefit correspondence theorem, health improving aspects of quality of care in domains such as physical, mental or social well being, under the WHO (1948) definition of health, have been suggested as potentially important, particularly in assessing performance in non-acute hospital activities. These perceived utility bearing aspects of care can be included as equivalent disutility events where there are either standards or cardinal measures of functioning. Utility bearing effects of quality of care can be simply reframed as disutility events with measures such as admissions where standards are not met, or average functional limitation across patients. As with other disutility events, these need to be verifiable (and hence auditable) measures of activity. In relation to measures of standards, adjustment for baseline predictive factors can be undertaken either on expected rates, or by adjusting levels of standards for patient covariates.

In satisfying correspondence theorem conditions quality of care can be incorporated in relative performance measurement consistent with maximising net benefit, while preventing incentives for cream-skimming and cost (and event)-shifting. Therefore, the net benefit correspondence theorem, applied at a clinical activity level in combination with decision analytic methods provides a systematic framework for robustly including quality of care in efficiency measurement consistent with maximising net benefit.

Chapter 6: Including quality of care in funding

6.1 Overview

Current public hospital funding mechanisms based on case-mix activity ignore differences in quality or outcomes of care, despite evidence of variation in practice. As a consequence, perverse economic incentives are created to reduce quality until the marginal cost from such reduction is zero. To create economic incentives for appropriate quality of care requires funding conditional on quality of care, with a trade-off between the cost and value of quality.

Measurement of relative performance consistent with maximising net benefit, as illustrated in chapter 5, was based on an ordinal one-to-one correspondence between maximsing net benefit and minimising cost plus value of disutility events. This correspondence is also demonstrated to be cardinal. Consequently, funding conditional on disutility event rates (mortality, morbidity, readmission), when scheduled at the decision maker's values for avoiding disutility events, is shown to be consistent with maximising net benefit per admission, under correspondence conditions.

A sequential two-stage funding mechanism is identified, and illustrated, that ensures budgetary constraints are maintained while allowing hospitals, their medical staff and administrators, to adapt to economic accountability for effects of quality of care. Conditioning payments on disutility event rates relative to best practice, initially at the current industry shadow price of avoiding disutility event rates, ensures scheduled payments remain within current case-mix payments. The scheduled value of quality can then be progressively move towards net benefit maximisation either until the decision maker's value is reached, or a second stage buffer payment[1], as the residual of current average cost case-mix funding, is exhausted.

Whether net benefit maximising value of quality can be reached for any given DRG at the level of current payments per admission, under this payment mechanism, is an empirical question of technical, versus allocative, inefficiency. Where a net benefit maximising value is not achieved within current case-mix payments per admission,

[1] Buffer payments would ideally be based on underlying as opposed to observed best practice, adjsuting for the effects of sample variation (as outlined in section 8.4).

economic incentives for quality of care are still more appropriate with a positive, as opposed to implicit 0, value for quality under current case-mix funding. At an industry level, the proposed mechanism in valuing effects of quality relative to best practice provides economic accountablilty for quality of care. Under correspondence conditions, incentives to 'quality skimp' are removed, as is the scope to hide technical inefficiency behind low quality of care (given for each clinical activity average industry quality lies above cost minimising levels, in practice).

As was the case in performance measurement a clincial activity level is suggested in allowing correspondence conditions to be satisfied in determining funding, given the ability to:

1. standardise for differences in expected disutility event rates and costs;
2. allow for multiple disutility events, with the linear nature of correspondence, including linking to effects post-separation;
3. reframe utilty-bearing aspects of qualty of care as disutility events, and;
4. verify effects of care with clinical audit and peer review.

6.2 Funding mechanisms ignoring quality of care

In activity based funding systems, such as prospective case-mix payments per admission for hospitals (Duckett, 1998), funding mechanisms operate as though technology is optimally and uniformly applied in maximising health gain for given cost of care. This assumption is made, despite evidence of practice variation, with high and variable rates of adverse events (e.g. mortality, morbidity, and readmission) in hospitals for individual clinical activities in the:

UK (NHS, 2000);
US (Marshall, Shekelle, Leatherman and Brook, 2000) and;
Australia (ACHCS, 2001; Fahey and Gibberd, 1995).

While calls have been made to use evidence of safety and quality of care in hospitals to improve application of evidence-based medicine (McKee, 1997; Gibberd, Pathmeswaran and Burtenshaw, 2000), differences in quality are ignored in these funding mechanisms. As a consequence, economic incentives are created by such mechanisms to reduce quality of care while the marginal costs of quality of care is

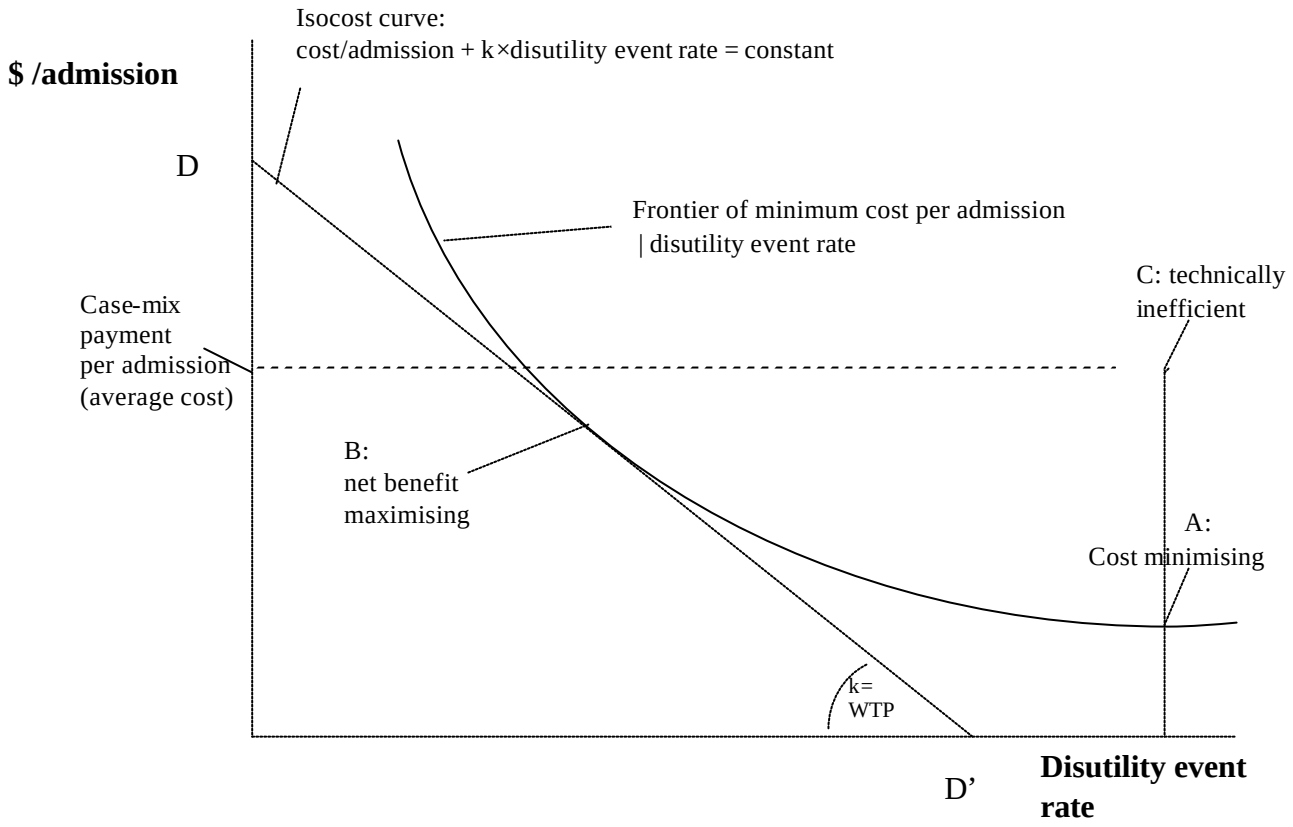
greater than zero, engaging in what Rice and Smith (2001) describe as 'quality skimping'.

For funding mechanisms to avoid such economic incentives, payments need to be conditioned on quality of care. Various ad-hoc subsidiary funding mechanisms, such as the 'rewards results initiative' in the US (NHCPI, 2002), have recently been proposed, conditioning a proportion of payments on quality standards. However, to enable an optimum mechanism with continuous (as opposed to local) appropriate economic incentives, a systematic method is required that provides an appropriate trade-off between the cost and value of quality of care. This chapter addresses how the framework provided by the net benefit correspondence theorem can be applied in funding to move from current case-mix funding towards a funding mechanism with economic incentives for net benefit maximising quality of care.

6.2.1 Problems with economic incentives created by case-mix funding

Case-mix payment mechanisms for any given DRG implicitly reflect historical average cost of care across a sample of hospitals, but do not reflect decision maker's values for the effects of quality of care. The lack of economic accountability for effects of care, described by case-mix proponents as 'clinical neutrality' (Brook, 2002) of case-mix funding, does not, however, imply neutral economic incentives. As figure 6.1 illustrates, using a funding mechanism that does not condition on quality of care creates economic incentives to reduce quality of care (increasing disutility event rates) to a point of cost minimisation, point A in figure 6.1.

Figure 6.1 Case-mix cost minimisation versus net benefit maximisation



At cost minimising quality of care at point A, intramarginal cost reductions from reducing quality of care across patients are just offset by expected additional inframarginal costs from treating additional complications within admission, resulting from lower quality of care. Quality of care per admission is implicitly valued at 0. By comparison, at net benefit maximising quality of care at point B, the cost of quality is traded off against the value of quality of care, where marginal cost of quality equals marginal value of quality (k=WTP to avoid disutility event).

Economic incentives for cost minimising quality of care, in funding and performance measurement under case-mix payments, may be mitigated to varying degrees by regulation of clinical practice and health-related arguments in providers' objective functions. At an industry level, the gatekeeper role of evidence-based medicine in health technology assessment, and processes of clinical monitoring, such as accreditation, peer review and associated clinical performance measurement, can each contribute to opposing economic incentives for quality of care minimising cost per admission. Depending on hospital, the relationship of hospital administrators with providers for any given clinical activity, arguments in providers' objective functions, such as those related to the hippocratic oath and professional standing, can also play a role in opposing economic incentives for cost-minimising quality of care in practice.

Holding hospitals economically accountability for costs of care alone allows scope for technical inefficiency to be hidden behind lower quality of care, as well as leading to allocative inefficiency in implicitly valuing quality of care at zero. Both of these inefficiencies lead leading to losses in net benefit. Technical inefficiency can be hidden under case-mix funding models where, for example a hospital such as C in figure 6.1 has average cost per admission but below average and cost minimising quality of care.

In the absence of economic accountability for quality of care in funding, for any given clinical activity, the presence of differing levels of quality across hospitals allows the scope for individual hospitals to effectively free-ride on industry quality of care above a cost minimising level. The extent of this scope to hide technical efficiency behind lower quality of care depends on degree of variability in quality. This is likely to be significant under case-mix funding, where payments for any clinical activity (DRG) reflect average expected costs across a sample of hospitals (Commonwealth

Department of Health and Aged Care, 2000) with variable quality of care (ACHCS, 2001; Fahey and Gibberd, 1995).

In summary, there is an active economic incentive for hospitals, in each clinical activity, to have cost minimising level of quality under case-mix funding. However, in practice, quality above that of cost minimisation under case-mix funding can be expected to the extent that positive values for health effects of care manage to oppose these economic incentives. Under case-mix funding for any given inpatient activity (DRG), where quality of care at an industry level is above a cost minimising level in practice, technical inefficiency can be hidden behind lower quality of care.

6.2.2 Problems of incentives for cost-shifting beyond admission

An economic incentive to reduce quality of care, while the expected marginal cost of such quality reduction is greater than 0, is perverse in considering the reduction in net benefit from the technical and allocative inefficiency this encourages. However, the extent of loss in net benefit only becomes truly apparent at a health system level. Lower quality of hospital care, in general, leads to a greater rate and complexity of follow-up health care, with the notable exception of where patients die within admission. As described in chapter 2, with incomplete vertical integration (Evans, 1981) cost-minimising quality of care within admission will not minimise costs of care in treating a patient population across time, where worse health outcomes at separation lead to increased requirements for care beyond-separation.

Even in comparison across the same DRG, an admission cannot be considered homogenous in value of treatment without accounting for the effects of care, including beyond-separation effects. Without considering effects of care, the lack of homogeneity as to what an admission constitutes also makes it unclear as to what is a cost-minimising level of care, and hence at what point economic incentives to reduce quality stop. To create incentives for appropriate (net benefit maximising) quality of care within the wider context of a health system, requires that effects beyond, as well as within, separation, are accounted for. Health effects of care should either be included with data linkage to effects beyond-separation or by health state at point of separation. Otherwise, incentives are created for cost-shifting and abrogating of the responsibility for patient care.

At some point, lower quality of care results in expected health losses relative to next best alternative forms of care. As section 4.6.2 outlined, prospect theory (Kahnemann and Tversky, 1978) predicts losses have greater disutility than utility from equivalent gains, with willingness to accept health losses suggested, empirically, to have a 2-3 fold greater value than willingness to pay for equivalent health gain (Willan, O'Brien and Leyva, 2001). This may be of particular concern, given high hospital reinfection, mortality and other iatrogenic event rates reported in hospitals within systems using case-mix funding (Marshall, 2000; ACHCS, 2001; Fahey and Gibberd, 1995).

It is a moot point for each clinical activity (DRG) as to what threshold rate of disutility events (quality of care) reflects an expected loss in health effects relative to a level of endowment from an entitlement to health care. However, where funding provides active economic incentives to lower quality of care and shift costs beyond point of separation, the risk of incremental health loss increases at both a system and individual hospital level. Failing to value, or account for, health effects in funding hospitals provides economic incentives for quality of care that may potentially result in expected net harm relative to a minimum threshold entitlement, or next best alternative forms of care (e.g. a general practice setting, or even doing nothing).

6.3 Finding a funding mechanism consistent with maximising net benefit

For a funding mechanism to align with maximising net benefit per admission, and to create appropriate incentives in trading off value and cost of care, payments conditional on quality of care should reflect value of care. However, if conditioning were based on expected costs for given effects then there would be no constraint on quality of care, potentially resulting in a spiralling in choice and use of more complex technology. Additionally, where expected costs increase at levels of quality lower than a cost minimising level, particularly perverse incentives can be created by payments based on expected costs. Critical appraisal of a study of relative performance in hospitals by Zuckermann et al. (1994), using stochastic frontier analysis, described in detail in section 8.4.3, highlights this problem. Perversely, expected costs were adjusted upwards for hospitals with standardized disutility event rates in both the lower and

upper decile. Such adjustment would clearly fail to create appropriate incentives for quality of care in funding mechanisms.

However, if payments reflect value, rather than expected costs of disutility event rates (standardized for patient risk factors and including effects beyond-care to reflect effects of quality of care), then hospitals face an appropriate trade-off between the cost and value of quality of care. Economic incentives for quality of care consistent with maximising net benefit are created if payments reflect differences in value of quality in a given patient population. A one-to-one cardinal correspondence between differences in net benefit and costs plus disutility event rates, valued at WTP as in net benefit, enables such a funding mechanism under correspondence conditions, with effects of quality of care represented by differences in disutility event rates.

6.3.1 Net Benefit Correspondence Theorem – absolute differences

A one-to-one cardinal correspondence exists between maximising net benefit per admission and the sum of cost per admission plus rates of disutility events valued by the decision maker in monetary terms as in net benefit, under the assumptions that:

1. *differences in effects of care can be captured by disutility events in patients and;*
2. *hospitals either face the same net benefit comparator, or differences between hospitals, in expected costs and disutility event rates are adjusted for.*

Proof: (multiple disutility events, common comparator case)

Let all combinations of disutility events be represented by (1,…,m) with associated event rates of these combinations (DU1,..,DUm) and values from avoiding these combinations of disutility events (k1,..,km). Then, under the first assumption, net benefit for any hospital (i) can be represented as:

$$\begin{aligned} NB_i &= k_1 \times (DU1_{comp} - DU1_i) + ... + k_m (DUm_{comp} - DUm_i) - (C_i - C_{comp}) \\ &= -(k_1 \times DU1_i + .. + k_m \times DUm_i + C_i) + (k_1 \times DU1_{comp} + .. + k_m \times DUm_{comp} - C_{comp}) \end{aligned} \quad (6.1)$$

Without loss of generalization, let

$$NB_i - NB_j = c,$$

where c is a constant value.

All comparator cost and effect terms cancel in comparing hospitals with the same comparator under the second assumption.

$\Leftrightarrow$

$$(k_1 \times DU1_j + .. + k_m \times DU_j + C_j) - (k_1 \times DU1_i + .. + k_m \times DUm_i + C_i) = c \qquad (6.2)$$

QED (multiple disutility events, common comparator case)

Proof: multiple disutility bearing events and differences in expected costs and effects

Let all potential combinations of disutility events observed in patients be represented by $(1,...m)$ with rates of these combinations across patient populations $(DU1,...,DUm)$ and associated values from avoiding these combinations of disutility events be $(k_1,...,k_m)$. Then under the first assumption of the correspondence theorem, net benefit for any hospital can be represented relative to a comparator, representing their expected costs and effects as:

$$NB_i = k_1 \times (DU1_{comp_i} - DU1_i) + ... + k_m (DUm_{comp_i} - DUm_i) - (C_i - C_{comp_i}) =$$

$$-(k_1 \times DU1_i + .. + k_m \times DUm_i + C_i) + (k_1 \times DU1_{comp_i} + .. + k_m \times DUm_{comp_i} - C_{comp_i}). \qquad (6.3)$$

Without loss of generalization (order is arbitrary in establishing a correspondence), let $NB_i > NB_j$.

Then from (6.1) $\Leftrightarrow$

$$-(k_1 \times DU1_i + .. + k_m \times DUm_i + C_i) > -(k_1 \times DU1_j + .. + k_m \times DUm_j + C_j) + z, \qquad (6.4)$$

where

$$z = k_1 \times (DU1_{comp_j} - DU1_{comp_i}) + .. + k_m \times (DUm_{comp_j} - DUm_{comp_i}) + C_{comp_j} - C_{comp_i}$$

Multiplying both sides of (6.4) by minus 1, the sign changes and we translate from maximising net benefit to minimising net loss per admission:

$$\Leftrightarrow \qquad C_i + DU1_i \times k_1 + ... + DUm_i \times k_m < C_j + DU1_j \times k_1 + ... + DUm_j \times k_m \text{-z}. \qquad (6.5)$$

Therefore, under the second assumption, if differences in expected costs and disutility events are adjusted for, there is a one-to-one correspondence with maximising net benefit.

QED (multiple events and adjusting for differences in patients expected costs and disutility event rates)

Under assumptions of the correspondence theorem, differences in minimising cost plus disutility events, valued at willingness to pay, has a cardinal correspondence with differences maximising net benefit, where hospitals face a common comparator. Where hospitals face patient populations with different expected costs and disutility event rates (that is, they do not face a common comparator), maintaining this cardinal correspondence requires that relative differences in expected cost and disutility events are adjusted for.

6.3.2 A funding mechanism with incentives for net benefit maximisation

Where assumptions of the correspondence theorem are satisfied, costs per admission of hospitals, adjusted for any differences in patient risks, are expected to be the same at any given level of quality. Absolute differences in the value of incremental effects are also equivalent to the value of incremental differences in disutility event rates, adjusted for any differences in risks of patient populations treated.

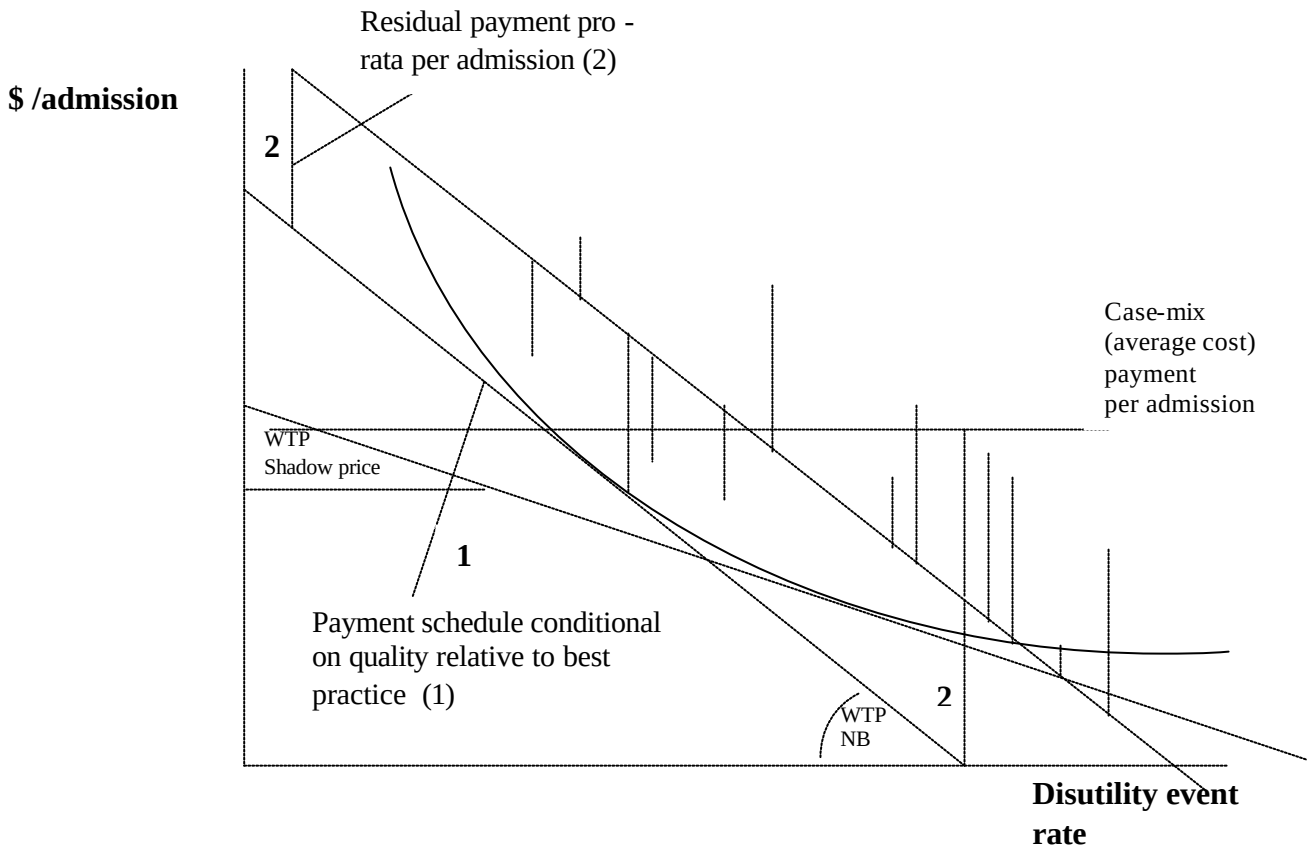
Net benefit is maximised at the point of tangency, between a frontier minimising cost and disutility events and an isocost curve whose slope represents the decision maker's value of avoiding disutility events (point B in figure 6.1). To create incentives for quality of care consistent with net benefit maximisation, reimbursement for any hospital can be made conditional on quality of care, relative to the cost for the net benefit maximising target on the frontier at the decision making value (k=WTP). These payments, conditional on disutility rate, are represented by the isocost curve with slope k (WTP) tangent to the frontier (DD' in figure 6.1).

For any disutility event rate below (quality above) the net benefit maximising level (B in figure 6.1), additional payments will not compensate for expected additional costs, given convexity of the best practice frontier. Similarly, for disutility event rates above the level at B (quality below), reduction in expected payments would more than offset expected reduction in costs of care. This payment mechanism, therefore, creates economic incentives for net benefit maximising quality of care.

For any given technology, estimated by observed best practice in DEA, this payment mechanism represents the lowest funding level payments that can be feasibly conditioned on, to provide net benefit maximising incentives. However, using DEA, a buffer payment to allow for sample variation in estimating the frontier of best practice (discussed in section 8.4), may be seen as necessary.

In general, whether an explicit amount to act as a buffer for sample variation is set or not, any remaining budget per admission for that activity under current case-mix payments can be distributed on a fixed amount per admission (figure 6.2).

Figure 6.2 A funding mechanism maximising net benefit



Whether any buffer payments per admission are possible, within current case-mix (average cost) funding per admission, is an empirical question for any given DRG, depending on the degree of technical, relative to allocative, inefficiency in production of net benefit. To the extent improvements in quality of care result in reduction in patient morbidity, the need for, use, and cost, of services beyond-separation are expected to reduce.[2] Any such cost savings from reduced treatment in a patient population over time, across the health system, could be seen as additional to that of case-mix funding per admission, in improving quality, while remaining budget neutral.[3] If at a given decision maker's WTP, the net benefit maximising peer's cost per admission is below average current payment levels, payments per admission made on the first, or scheduled, part of the funding mechanism will remain within case-mix industry budget per admission.

In moving from a case-mix payment system, with implicit economic incentives for a 0 value of quality, it is predictable that the shadow price of quality will lie below the desired value and that technical inefficiency, hidden by low quality, will be present. Payments conditional on disutility event rates, representing effects of care under correspondence conditions, provide the accountability and exact incentives required to convert technical inefficiency into quality improvement, and move the shadow price towards a net benefit maximising value.

In practice, as with case-mix payments, to avoid incentives for cream-skimming and cost (and event) shifting, the schedule requires differences in baseline risk and effects post-separation to be adjusted for. Adjusting for these pre- and post-care effects are necessary to satisfy correspondence theorem assumptions of a common comparator and coverage of effects by disutility events, as discussed in section 5.7 for performance measurement. Standardising for baseline predictive factors in funding is considered in section 6.8.

---

[2] While cost per admission is not proposed to change from that of current case-mix funding, higher quality of care may result in lower costs to the health system in treating a patient population over time. The potential exception to this is the notable case where higher quality reflects lower mortality within admission.

[3] However, the potential for demand to increase with higher quality of care still suggest a role for capitation payments, discussed in detail in section 7.4.3.

6.3.3 Payment schedules providing incentives to maximise net benefit

In providing incentives to maximise net benefit per admission, the proposed schedule of payments per admission for any hospital (i), conditional on each hospital's disutility event rate and the cost per admission of the best peer at decision maker's value of avoiding disutility events ($k = WTP$), can be expressed as:

$$payment_i = C_{peer(k)} + k \times (DU_{peer(k)} - DU_i) + A \qquad (6.3),$$

Where:

$C_{peer(k)}$ is cost per admission of the best practice peer at decision maker's value of avoiding disutility events;

k is the decision maker's value of avoiding a disutility event;

$DU_{peer(k)}$ is the disutility event rate of the peer maximising net benefit at k;

DU_i is the disutility event rate for hospital i, and;

A is a residual buffer payment per admission.

It is important to understand that, unlike conditioning on quality of care, the proposed funding mechanism in trading off cost and value of quality prevents spiralling of quality beyond an appropriate level.[4] Relative to payments based on expected costs per admission conditional on quality of care[5], the proposed payment mechanism punishes those with quality of care either below, or above, a net benefit maximising level. Relative to case-mix payments, hospitals with quality of care above industry average (disutility event rates below industry average) are rewarded, while those hospitals with below average quality are punished.

To allow continuity with case-mix funding, the buffer can represent the residual of case-mix funding per admission once payments conditional on quality of care, in the first part schedule, have been made.[6] For example, if the proposed budget per

[4] Case-mix payments provide economic incentives for cost-minimising quality of care. It is therefore unlikely that hospitals have a quality level above (disutility event rate below) that of net benefit maximisation. Under the proposed funding mechanism, where correspondence conditions are satisfied, economic incentives to increase quality are exhausted where the expected marginal cost of increasing quality becomes greater than k.

[5] Section 9.3.4 examines problems associated with funding based on expected costs in critically appraising the SFA approach of Zuckermann et al. (1994).

[6] A minimum buffer payment would ideally allow for sample variation where the frontier is estimated with DEA. Future research in use of Bayesian shrinkage estimation methods with DEA or SFA methods is suggested in section 9.3.

admission were equivalent to case-mix payments, in the year of performance measurement, say C_{ave}, then the constant residual payment is equivalent to:

$$A = C_{ave} - C_{peer(k)} + WTP \times (DU_{ave} - DU_{peer(k)}), \qquad (6.4)$$

where:

$C_{peer(WTP)}$ is the cost per admission of the best practice peer at the threshold value of WTP;

DU_{ave} is the average disutility across hospitals and;

$DU_{peer(k)}$ is the disutility event rate for the best practice peer at the decision-maker's value of avoiding disutility events, *k*.

The total payment per admission for a given hospital (*i*) with associated disutility event rate (DU_i) would then be given by:

$$payment_i = C_{ave} + k \times (DU_{ave} - DU_i) \qquad (6.5)$$

In moving to a funding system conditioning on quality of care, the payment mechanism in (6.5) changes the distribution, but not the average industry payment per admission, from that of case-mix payment. Relative to case-mix payments, negative budgetary impacts will, appropriately, be greatest for hospitals that have hidden technical inefficiency behind low quality of care. In contrast, with incentives to minimise costs per admission, evidence-based medicine is supported, individuals and social institution do not have to defend appropriate quality against perverse economic incentives, and the risk of creating incentives for quality of care representing expected health losses from care is avoided.

Whether the net benefit maximising value of quality can be reached in first stage scheduling, within current case-mix payment per admission, is an empirical question for each DRG. In a static world, this theoretically depends on whether case-mix payments currently are greater than the underlying best practice cost per admission, at the net benefit maximising value for quality of care.

In Australia, case-mix payments reflect average cost per admission, as relative weights are calculated from average costs in the National Hospital Cost Data Collection (NHCDC) sample of hospitals (Commonwealth Department of Health and Aged

Care, 2000). Where case-mix payments are made from a globally capped, capitation-based funding system, total inpatient activity levels can be used to recalibrate payments per case-mix weighted admission. It is this mechanism that allows current incentives to remain within budget. However, while aimed at improving technical efficiency, this mechanism, in failing to account for quality of care, provides economic incentives to lower quality of care (allocative inefficiency), and the ability to hide technical inefficiency behind lower quality of care.

In general, the question of whether net benefit maximising funding is feasible, within current case-mix funding per admission at an industry level, can be seen as one of whether average costs of current practice are above that of best practice, conditional on the decision maker's value of quality. This, in turn, is a function of the degree of technical inefficiency, versus allocative inefficiency, at the decision makers' value for quality. At an industry level, if inefficiency is attributable to technical, rather than allocative, inefficiency, then the necessary condition exists for net benefit maximisation to be reached. In practice, in a dynamic world, whether net benefit maximising incentives from funding are able to be achieved within current case-mix funding levels, will also depend on shift factors such as; patient risk factors; prices of inputs over time; technology changes over time, but also; the ability to change behaviour of hospitals across the industry with appropriate economic incentives.

6.4 Changing behaviour to match incentives

A transitional problem in undertaking a change in one step from funding independent of quality, to funding based on decision makers' value of quality of care, is that some hospitals may not be able to plan and adapt to accountability for quality of care fast enough. Hospitals need to move from practices based around economic accountability for costs within admission, to accountability for costs and quality of care and within the context of effects on the health system, assuming correspondence conditions are satisfied. Appropriately, this adjustment is likely to be hardest for hospitals which have hidden inefficiency behind low quality of care under case-mix funding. For policy makers, concerns also arise that the funding mechanism can ensure budgetary control.

To ensure budget constraints are maintained, and a smooth cultural change from activity-based funding with implicit 0 value of quality, a sequential mechanism may be considered. While the ideal is funding conditional on disutility event rates (quality of care), at a WTP value maximising net benefit, in practice this goal may only be achieved with a schedule that allows hospitals to plan, and have ownership over, change in strategies and culture.

In considering hospital cultural change, it is particularly important to be cognizant of the framework hospital payment mechanisms provide for the internal relationship between administrators and clinicians (Harris, 1977). For hospital administrators, the change in framework requires moving from an accounting role in minimising costs to considering opportunity costs for health outcomes in coordinating and supplying resources. Administrators need to account for trade-offs between costs and effects of quality of care in facing challenges of how to organize and motivate resources to maximise net benefit per admission. For clinicians, the change from a 0 to positive value of quality should more naturally align with other arguments in their objective functions in acting as agents for patients, while making them economically accountable for these outcomes of care.

In general, by providing a trade-off between value and cost of quality of care, consistent with maximising net benefit, a framework for more appropriate roles and incentives is created. The tension between hospital administrators attempting to minimise costs per admission and clinicians to maximise health outcomes per admission is diminished, with scheduled values providing guidance and signals from policy makers. Policy implications, and effects on internal hospital negotiations in moving to a framework of performance measurement and funding based on the correspondence theorem relative to that of current case-mix funding, are explored in detail in section 7.3.

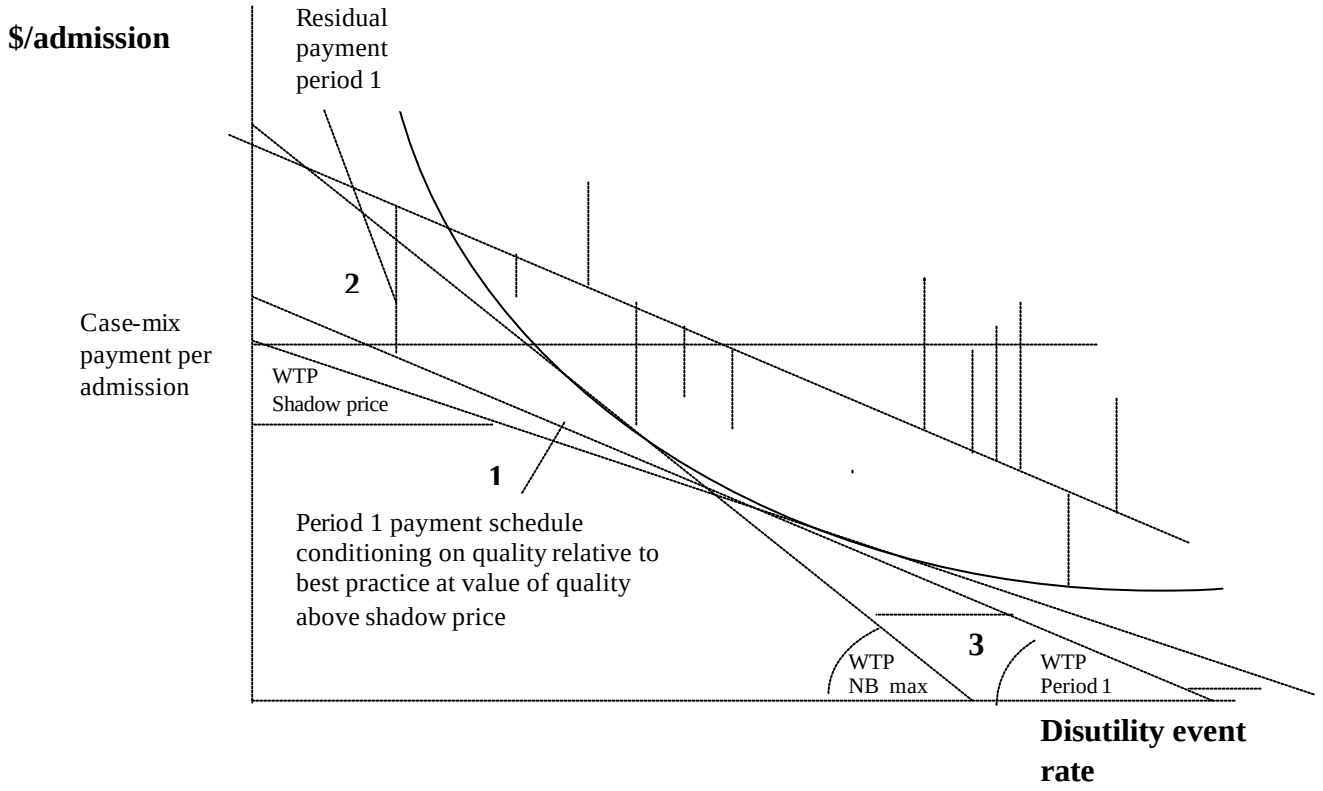
6.4.1 A sequential funding mechanism

In transition from cost minimisation under case-mix funding towards net benefit maximisation in conditioning payments on effects of care, a conservative value for avoiding disutility events could be initially employed. To ensure budget constraints per admission are maintained, the current industry shadow price for avoiding disutility

events, identified using the method described in 5.6.5, could be adopted in first stage payments conditional on effects of care, relative to best observed practice at this value. This would at least move from an implicit 0 value of quality under case-mix funding, to rewarding current behavior in valuing quality[7]. Over time, the schedule could then move towards decision makers' value for disutility events (quality of care) in an incremental and planned manner, with increasingly appropriate incentives for net benefit maximisation, as illustrated in figure 6.3.

[7] This shadow price is, by definition, positive under constant returns to scale and strong disposability of cost and disutility events as inputs and admissions as output, given consequent convexity of the unit isoquant in the cost-disutility plane. With economic incentives under case-mix funding for a 0 value for quality, the shadow price should, however, lie below the decision maker's value maximising net benefit.

Figure 6.3: A sequential funding mechanism: moving towards net benefit maximisation within a budget per admission



In figure 6.3 a sequential two-part funding mechanism (figure 6.3) is illustrated, where:

1. First-stage payments are conditional on disutility event rates relative to the peer, or target, at the value of current shadow price in industry behaviour (assumed positive);
2. In the second stage the remainder of budget is apportioned per admission and;
3. In subsequent periods the schedule shifts the value of quality (avoiding DU events) towards that of maximising net benefit per admission.

For any given DRG, the process is proposed to continue until either:

(i) a steady state is reached, where the value of quality inherent in the schedule provides incentives for maximising net benefit per admission, or;

(ii) the budgetary constraint has been reached in first stage payment (where the buffer falls to 0 or some predefined level, allowing for sample variation).

6.4.2 Transition state incentives

In sequentially increasing the scheduled value of quality towards a net benefit maximising level, incentives are created to increase quality to where expected costs of such quality improvement match those of the scheduled value. Under correspondence conditions, increasingly appropriate incentives for quality of care are, therefore, created at each stage, in accounting for effects of care.

Across the industry, gains from appropriate incentives in reducing technical and allocative inefficiency are translated into improving quality. Given the shift is from a system which implicitly valued quality at 0, and provided scope for technical inefficiency to be hidden by low quality of care, the focus on quality improvement is appropriate. The emphasis on increasing quality is particularly appropriate for clinical activities (DRGs), where economic incentives to lower quality to cost minimising levels may have resulted in care with expected health losses, relative to next best care at an individual hospital, or even an industry, level.

In undertaking a transition towards a system of funding with appropriate incentives for quality, the sequential two-part funding mechanism proposed:

1. allows budgetary constraints to be maintained

2. provides achievable and incrementally more appropriate incentives for quality improvement at each stage and;
3. is conducive to planning of administrators.

Compared with a one-step shock, this should enable a culture change in moving to incorporating value, and in providing accountability for quality of care.

6.4.3 Final (steady state) incentives

In a final (steady state), if the funding schedule is based on the decision maker's value of quality care, then under correspondence theorem conditions, hospitals have an economic incentive to maximise net benefit. Given diminishing marginal returns to increasing quality inherent in the convex nature of the constructed frontier, expected marginal funding from increasing quality of care is less than expected marginal cost above net benefit maximising level, and greater than expected marginal cost below this level.

Whether net benefit maximisation is reached within current budgets or not, in creating accountability for effects, as well as costs of care, economic incentives for quality of care under correspondence conditions will be more appropriate. If the final scheduled value is below that of the decision maker, then incentives for quality of care will still have moved towards net benefit maximisation from current incentives for cost-minimisation. Incentives for allocative inefficiency in valuing quality, and the scope for hiding technical inefficiency, would be removed and, *ceteris paribus* performance of hospitals should improve, with higher quality of care for given resources.

6.5 Illustrating funding conditional on quality of care

The proposed funding mechanism, conditioning payments on disutility event rates (quality of care), is illustrated for DRG e62a (respiratory infection) across 45 Australian public hospitals form New South Wales in 1998-99, with higher quality of care represented by lower mortality rate. These are the same hospitals and data used to illustrate relative performance measurement under output specifications of quality, in chapter 4, and an input specification under the correspondence theorem, in chapter 5.

As was the case in performance measurement in chapter 5, for purposes of illustrating the proposed funding mechanism, to satisfy correspondence conditions, the assumption needs to be made that there are no differences across compared hospitals in: average expected costs and disutility event rates for patient populations treated within DRG, or; beyond-separation effects of care. In reality, any differences in prior risk and post-separation effects would need to be adjusted for, to prevent incentives for cream-skimming and cost- and event-shifting.

The absence of adjustment for patient risk factors and data linkage to effects beyond-care are limitations, in illustrating the application of the correspondence theorem. However, these limitations relate to available data in undertaking this thesis, rather than the method described, or existence of data more generally, and are made clear due to the explicit nature of correspondence theorem assumptions.

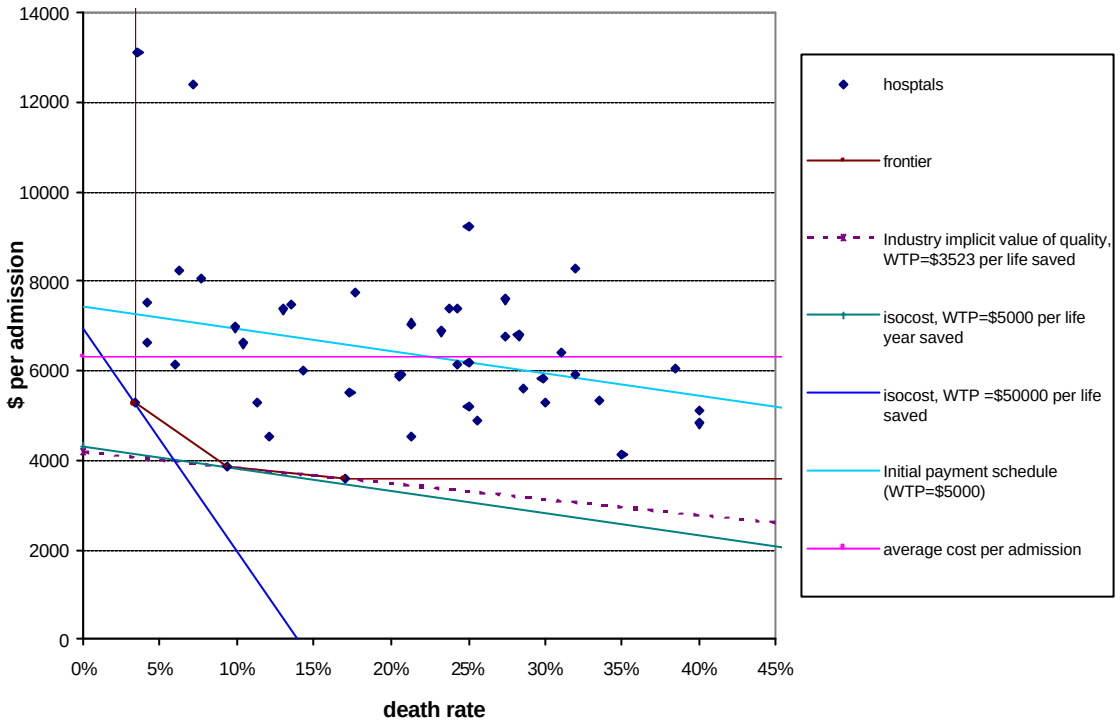
At a clinical activity level, the ability to employ decision analytic methods allows comprehensive, evidence-based identification of effects of care including those beyond-separation (as described in section 5.7.2), and adjustment for patient risk factors within DRG (as described in section 5.7.1). The linear nature of the correspondence theorem allows flexible inclusion of multiple disutility events and costs, including those beyond-separations. Relative utility-improving aspects from care can be reframed as disutility events where they can be expressed as meeting a standard of care, functional ability on a cardinal scale, or utility directly. Data linkage to effects post-care and patient level data, to model risk factors for disutility events between hospitals was not, however, available in this thesis.

The correspondence theorem challenges policy makers to undertake appropriate data linkage and risk factor modelling, informed by decision analytic methods, in satisfying correspondence conditions, and in allowing incentives for evidence-based medicine in practice. These policy challenges and the data requirement to meet them are explored in detail in section 7.2.

An implicit industry value (shadow price) of avoiding deaths in 1998-99 was estimated to be $3,523 in section 5.6.4, where industry economic efficiency as the cost-share weighted economic efficiency of hospitals was maximised (as illustrated in figure

5.6.2). To ensure that payments from first stage funding schedules (conditional on disutility event rates relative to best practice), are within the current case-mix budget per admission, this shadow price could be used as the schedule value in the initial period. However, given significant technical inefficiency, reported in table 5.6.3, the value of quality implicit in the schedule in the first period could be based at a value above current industry shadow price, at say $5,000 per death averted, while remaining within budget in first stage payments. This is illustrated in figure 6.4.

Figure 6.4 Intitial payment schedule for DRG E62a conditioning on quality of care and WTP=$5000 per life saved



At a scheduled value of \$5,000 per death averted, hospital 17 is the peer (economic efficiency 1), with a cost per admission of \$3,858 and a mortality rate of 9.4%. The payment schedule per admission for each hospital (i=1... 45) as proposed in equation (6.3) would therefore take the form:

$$payment_i = 3858 + 5000 \times (0.094 - DU_i) + A \qquad (6.6)$$

The residual or buffer payment per admission (A) can be estimated from equation (6.4), assuming the same disutility rate and same average industry budget per admission as in 1998-99 as:

$$A = 6332 - 3858 + 5000 \times (0.224 - 0.094) = \$3125 \qquad (6.7)$$

In figure 6.4, the vertical distance between the actual payment schedule and the first stage payment schedule (represented by the isocost curve at \$5,000 per life saved tangent to the frontier) represents this second stage fixed payment per admission of \$3125 as in (6.7).

Substituting A into equation (6.6), payments under the schedule with shadow price \$5,000 and budget equivalent to average expenditure for hospital (i=1... 45) would then be of the form:

$$\begin{aligned} payment_i &= 3858 + 5000 \times (0.094 - DU_i) + 3125 \\ &= 7453 - 5000 \times DU_i \end{aligned} \qquad (6.8)$$

Table 6.5.1 shows what the level of payments under this schedule would be, for each of the 45 hospitals at their current mortality rate, and how this compares with case-mix funding based on average cost per admission of \$6332 per admission.

Table 6.5.1 Funding schedule conditioning on quality of care within a budget (k=$5000 per life saved) for DRG E62a

Hospital	Mortality rate (%)	Payments/admission Conditioning on quality (k=$5,000/ life saved)	% change relative to average cost payment of $6,332 per admission
1	40.0%	$5,453	-14%
2	25.0%	$6,203	-2%
3	7.7%	$7,068	12%
4	7.1%	$7,095	12%
5	40.0%	$5,453	-14%
6	6.3%	$7,140	13%
7	35.0%	$5,703	-10%
8	14.3%	$6,738	6%
9	13.0%	$6,800	7%
10	4.2%	$7,244	14%
11	4.2%	$7,244	14%
12	32.0%	$5,853	-8%
13	38.5%	$5,529	-13%
14	3.6%	$7,274	15%
15	10.3%	$6,935	10%
16	25.0%	$6,203	-2%
17	9.4%	$6,984	10%
18	24.2%	$6,240	-1%
19	12.1%	$6,846	8%
20	24.3%	$6,236	-2%
21	13.5%	$6,777	7%
22	25.6%	$6,170	-3%
23	20.5%	$6,427	2%
24	30.0%	$5,953	-6%
25	21.3%	$6,389	1%
26	17.0%	$6,603	4%
27	6.0%	$7,154	13%
28	17.6%	$6,570	4%
29	11.3%	$6,889	9%
30	32.0%	$5,853	-8%
31	17.3%	$6,586	4%
32	27.4%	$6,083	-4%
33	3.3%	$7,286	15%
34	9.9%	$6,958	10%
35	23.8%	$6,264	-1%
36	25.0%	$6,203	-2%
37	29.8%	$5,961	-6%
38	23.3%	$6,289	-1%
39	31.0%	$5,902	-7%
40	20.6%	$6,423	1%
41	28.6%	$6,024	-5%
42	21.3%	$6,389	1%
43	33.5%	$5,775	-9%
44	27.4%	$6,084	-4%
45	28.3%	$6,039	-5%
total	22.4%	$6,332	0%

In reality, hospital quality of care should change with incentives inherent in the schedule. Appropriately, hospitals with the lowest quality of care (highest standardised disutility rate) have greatest incentive to increase quality. If quality does not improve, hospitals with low quality of care face budget shortfalls for that DRG compared with payments under case-mix funding. Hospitals that have hidden technical inefficiency, behind lower quality of care, face budget shortfalls relative to cost of care, as well as relative to DRG payments, if they do not improve quality of care.

In our case example, assuming mortality rates were standardised for baseline risk at admission between hospitals, and included beyond-care effects, then hospitals 1, 5, 7, 13 are suggested to be hiding technical inefficiency behind low quality care. If they do not improve quality of care (reduce disutility event rates), they face budget shortfalls, both relative to previous payments and relative to costs of care, given current quality. Hospitals with relatively high quality care are rewarded compared to current payments, but not as much as they should be to encourage net benefit optimising behaviour.

To allow a movement to incentives for maximising net benefit per admission, the value of quality implied in the schedule can be sequentially increased towards a desired value for quality of say $50,000 per death avoided, at a rate of $5,000 per period. The peer at a value of $50,000 is hospital 33 with a cost per admission of $5,283 and mortality rate of 3.3%.

A schedule based on $50,000 per life year saved (represented by the isocost curve tangent to the frontier with value of $50,000 per life year) is also represented in figure 6.4 as indicative of why a move to net benefit maximisation is likely to be unmanageable in one step. The residual payment, again (and this time, less believably) assuming no behavioural change from equation (6.3), is:

$$A = 6332 - 5283 + 50000 \times (0.224 - 0.033) = \$10{,}749 \qquad (6.9)$$

Consequently, the payment schedule from equation (6.4) is of the form:

$$payment_i = 16032 - 50000 \times (DU_i - 0.033) \qquad (6.10)$$

Assuming there is no industry change in mortality rates, for mortality rates above 35% the scheduled payments per admission become negative.

If, as predicted, there is behavioural change across the industry to increase quality, then the second-stage payment is reduced from $10,749, reducing to 0 if the mortality rate falls to 2.0% across the industry. The point at which any individual hospital's payment falls to 0 can be approximated (from equation 6.5), assuming the hospital does not affect industry average disutility rate as:

$$\begin{aligned} 0 &= C_{budget} + k \times (DU_{ave} - DU_i) \Leftrightarrow \\ DU_i &= DU_{ave} + C_{budget} / k \end{aligned} \qquad (6.11)$$

where:

DU_{ave} is the average disutility rate across hospitals and;

C_{budget} is the proposed budget per admission (case-mix funding in 6.5).

Assuming a proposed budget per admission reflects average costs (as in case-mix payments), then, under equation 6.11, the payments fall to 0, at a disutility rate of:

$$DU_{ave} + C_{budget} / k = DU_{ave} + 6332 / 50000 = DU_{ave} + 12.6\% = 35.0\% \qquad (6.12)$$

More generally, for payments to not fall below any minimum level per admission ($C_{\min}$ say), and again assuming the hospital is small enough not to affect the industry average, a hospital would need to have a disutility event rate below:

$$DU_{ave} + (C_{budget} - C_{\min}) / k \qquad (6.13)$$

In the case of a proposed budget based on current case-mix payments (average cost), the second stage buffer payment ensures payments are independent of the best practice peer, dependent only on overall budget per admission and the value of divergence from industry quality at the decision maker's WTP. Equation 6.13 makes clear that the room for divergence from industry average quality for any given clinical activity (standardised disutility event rate) lessens with increases in the value of avoiding disutility events (the decision maker's value of quality).

A sequential approach in shifting from case-mix payments (with an implicit 0 value for quality) towards a net benefit maximising value, may therefore be preferred, in allowing behaviour to change. In our case example, assuming correspondence

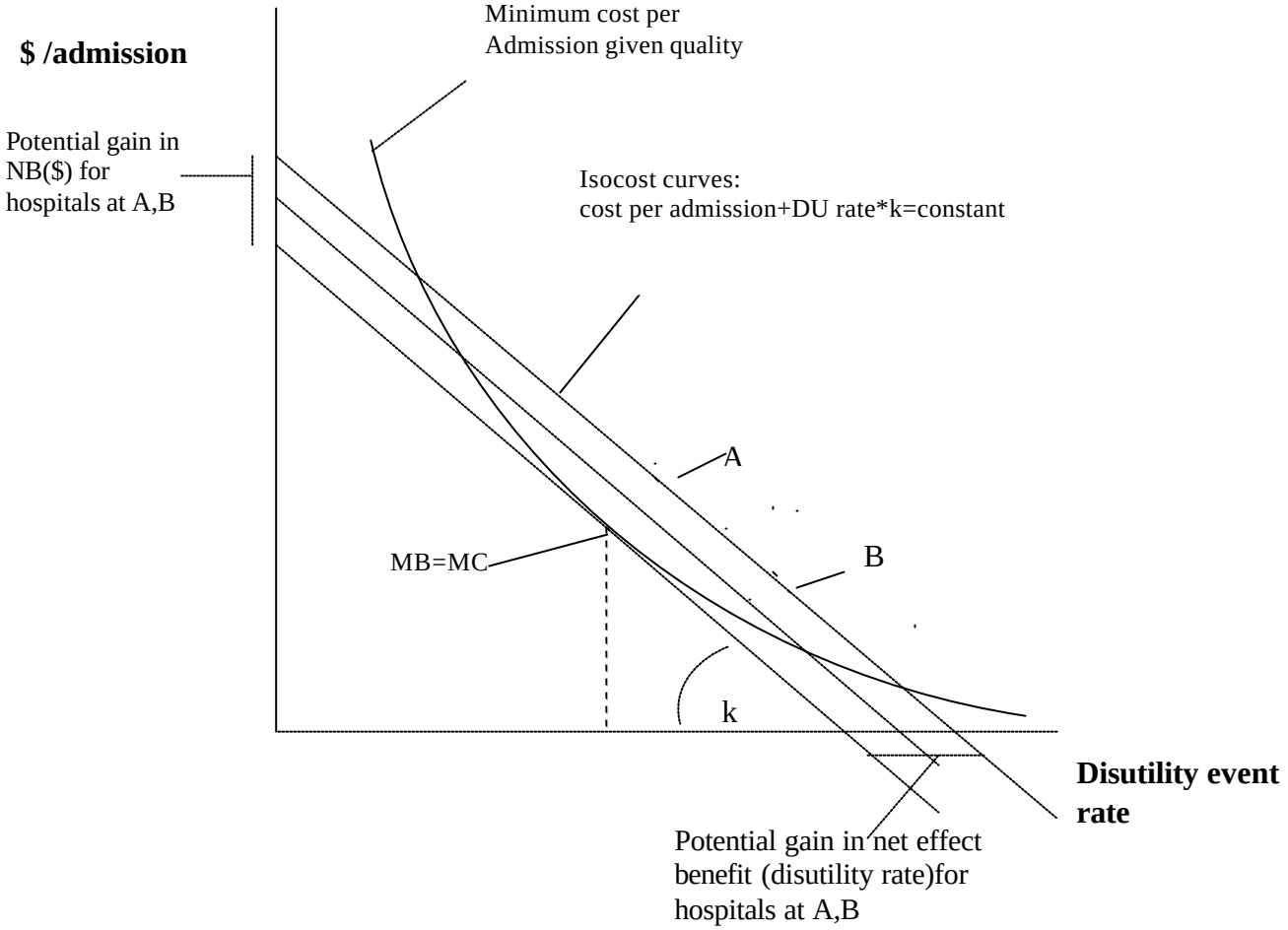
conditions were satisfied, immediately moving to a schedule with a value of quality of $50,000 per life year saved could create problems of adjustment, particularly for hospitals with very high disutility event rates, as illustrated in diagram 6.4 and in equation 6.13.

A sequential approach allows those with low quality (high rate of disutility event rates adjusted for pre-admission risk factors and post-separation effects of care) to catch up in improving quality of care and to reduce technical inefficiencies which had been allowed, with a lack of economic accountability for effects of care, under case-mix payment systems. However, the trade-off is that a slower movement towards net benefit maximising scheduled values and incentives sacrifices potential gains in net benefit from changing behaviour faster. In considering the optimum rate at which the value of effects might change, policy makers need to consider how fast hospitals are able to change practices entrenched under case-mix funding and adapt to economic accountability for quality of care.

6.6 Potential gains in net benefit

Under correspondence conditions, potential gain in net benefit per admission, for any hospital, can be estimated as the distance between the isocost curve on which they lie and that tangent to the best practice frontier (a unit isoquant under constant returns to scale in figure 6.5). In the case of DEA, this assumes observed best practice represents technology.

Figure 6.5: Potential gain in net benefit per admission



Gains in net benefit can be considered on the vertical axis in monetary terms, per admission, or on the horizontal axis as the disutility rate avoidable. The potential gain from acting in an economically efficient manner, at point of best practice, can be decomposed into technical efficiency (of net benefit) in radial contraction to the frontier, and allocative efficiency as a residual of technical and economic efficiency. At an industry level, the potential gains in net benefit per admission for each hospital can be multiplied out by number of admissions, and then summed across hospitals, to find industry potential gain in net benefit.

In our case example, if all hospitals were economically efficient (at an assumed value of k= $50,000 per life saved), then industry mortality could be reduced in absolute terms by 19.1 % and cost per admission by $1049, to the 3.3% mortality rate and $5,283 per admission of hospital 33. Across 2846 admissions, in 45 hospitals in 1998-99, this translates to potential savings of $3.0 million and 551 deaths averted, if industry behaviour mirrored that of best practice.

At a WTP of $50,000 per life saved, the potential gain in net monetary benefit is estimated at $10,749 per admission across the industry[8], or in terms of effect, equivalent to an absolute mortality reduction of 21.5%. Across 2846 admissions this translates to a net monetary benefit of $30.6 million, or net effectiveness benefit of 611 lives saved. Table 6.6.1 provides estimates for each hospital of the potential gain in net monetary benefit, at a value of quality of $50,000 per life saved, attributable to overcoming economic inefficiency. The potential gain is equivalent to the distance on the vertical axis between isocost curves on which hospitals lie, relative to the isocost curve tangent to the best practice frontier (hospital 33 at a value of $50,000 per life saved).

For each hospital, and across the industry, estimates of potential gain in net monetary benefit from economic inefficiency are also decomposed, in estimating that attributable to overcoming allocative inefficiency and technical inefficiency under constant returns to scale (Charnes, Cooper and Rhodes, 1978), technical inefficiency under variable returns to scale (Banker, Charnes and Cooper, 1984)), and scale inefficiency.

[8] Equivalent to the residual at $50,000 per life year saved in equation 6.9.

Table 6.6.1: Estimated potential gain in net monetary benefit (NMB) per admission ($), by source of inefficiency, at a WTP of $50,000 per life saved

Hospital	Technical inefficiency variable returns to scale	Scale Inefficiency	Technical inefficiency constant returns to scale	Allocative inefficiency	Potential gain in NMB Economic Inefficiency
1	$0	$6,456	$6,456	$11,422	$17,878
2	$5,648	$7,169	$12,817	$1,955	$14,772
3	$0	$4,642	$4,642	$357	$4,999
4	$0	$8,470	$8,470	$639	$9,109
5	$4,020	$3,517	$7,537	$10,552	$18,089
6	$0	$4,322	$4,322	$114	$4,436
7	$433	$2,380	$2,813	$11,901	$14,714
8	$2,366	$2,234	$4,600	$1,577	$6,177
9	$4,449	$1,390	$5,840	$1,112	$6,952
10	$0	$1,746	$1,746	$0	$1,746
11	$0	$1,926	$1,926	$770	$2,696
12	$12,637	$972	$13,609	$3,645	$17,254
13	$9,102	$1,264	$10,366	$8,090	$18,456
14	$0	$1,044	$1,044	$6,860	$7,904
15	$3,183	$707	$3,890	$943	$4,833
16	$7,106	$561	$7,667	$4,114	$11,780
17	$0	$0	$0	$1,624	$1,624
18	$9,375	$195	$9,571	$2,930	$12,501
19	$1,587	$0	$1,587	$2,010	$3,597
20	$7,136	$183	$7,318	$4,025	$11,344
21	$6,124	$0	$6,124	$1,139	$7,263
22	$4,248	$354	$4,602	$6,194	$10,796
23	$5,974	$0	$5,974	$3,229	$9,203
24	$6,089	$406	$6,495	$6,901	$13,395
25	$3,036	$152	$3,188	$5,010	$8,198
26	$0	$0	$0	$5,074	$5,074
27	$1,732	$91	$1,823	$365	$2,188
28	$6,958	$1,160	$8,118	$8,449	$16,568
29	$1,422	$1,203	$2,625	$1,312	$3,937
30	$5,699	$2,850	$8,549	$6,357	$14,906
31	$2,270	$2,270	$4,539	$2,695	$7,234
32	$6,141	$3,275	$9,416	$4,094	$13,510
33	$0	$0	$0	$0	$0
34	$2,981	$1,192	$4,173	$835	$5,007
35	$5,786	$3,665	$9,451	$2,893	$12,344
36	$354	$5,130	$5,484	$5,307	$10,790
37	$2,488	$5,390	$7,878	$5,805	$13,683
38	$3,890	$4,631	$8,521	$2,964	$11,485
39	$3,728	$5,921	$9,648	$5,263	$14,911
40	$0	$6,000	$6,000	$3,243	$9,243
41	$597	$6,568	$7,165	$5,772	$12,937
42	$354	$7,786	$8,140	$2,654	$10,794
43	$0	$7,292	$7,292	$7,955	$15,248
44	$0	$10,859	$10,859	$3,407	$14,266
45	$0	$9,627	$9,627	$4,395	$14,021
Industry ($)	**$2,405**	**$4,321**	**$6,726**	**$4,023**	**$10,749**

Of the potential gain in net benefit at an industry level, 63% is attributable to technical inefficiency under constant returns to scale and a residual 37% to allocative inefficiency. Under the assumption of variable returns to scale, the 63% attributable to technical inefficiency under constant returns to scale can be further decomposed to 22% attributable to technical inefficiency under variable returns to scale, and a residual 41% to scale inefficiency. As stressed in appendix 4.1, scale inefficiency should be seen as indicative, given the requirement for observed best practice to represent technical efficiency across the scale of production, for these estimates to be robust.

The larger proportion of inefficiency attributable to technical, rather than allocative, inefficiency, at $50,000 per life saved, is suggestive that net benefit maximisation can be achieved without additional funding per admission. This is reflected in the net benefit maximising peer hospital having a lower cost per admission than the industry average. However, there are limitations to this inference, both related to satisfaction of correspondence conditions, and whether the frontier estimated with DEA represents underlying best practice.

6.6.1 Limitations of inferences related to potential gain in net benefit

A practical interpretation of the assumptions made in estimating the position of the frontier and potential gains in net benefit, at a given value of avoiding disutility event rates, is that the minimum cost curve conditional on disutility event rate (quality of care) is that able to be reached by each hospital, if technically efficient. This implies that hospitals compared face the same expected cost (already required in case-mix funding) and baseline risk of disutility events in treating these patient populations, but additionally that the frontier represents underlying technology. Violations of these assumptions create limitations in interpreting potential gains in net benefit. At a clinical activity (DRG) level, differences in expected disutility event rates are able to be adjusted using logistic regression, given data for patient risk-factors as described in 5.7.1, while effects post-care can be simply included with data linkage as described in 5.7.2. Using DEA identification of comparators can also allow for scale effects using variable returns to scale formulations (Banker, Charnes and Cooper, 1984), while peer grouping can be used to overcome differences in objectives or technology available. However, how to allow for sampling variation, given uncertainty of effects and costs of care, is a more vexed issue, as considered in section 9.3 as part of future research.

6.7 Robustness of correspondence in funding

The correspondence theorem makes the two explicit assumptions of a common comparator and disutility events representing effects of care. As in a net benefit framework, the implicit assumption of a constant value of quality (willingness to pay to avoid events) is also made. The robustness of these assumptions, with respect to relative performance, has been explored in detail in section 5.7, and, in most respects, this discussion translates to a funding setting.

In brief, in capturing the effects of quality of care, not all disutility bearing effects of care will be observed within hospitalisation. This is particularly important to account for, given the discretion of providers with respect to point of separation, and incentives for cost- and effect-shifting in any fee-for-service based system. To mitigate against these perverse incentives requires the ability to cover effects of quality of care with disutility event rates, including utility improving aspects of quality of care, as well as linkage to post-separation effects attributable to quality of care.

Utility bearing aspects of care can be included wherever verifiable standards of functioning, or measures of functional limitation on a cardinal scale of disutility, exist. At a clinical activity (DRG) level of funding, monitoring and audit processes in preventing finessing of reported rates of such effects are facilitated using processes of clinical peer review. The attribution of post-care effects can be undertaken with data linkage informed by decision analytic methods, in both identification of effects and degree of attribution. Degree of attribution can be specific to post-care effect (e.g. specific types of readmission) and time dependent to allow for likelihood of external cause (weights diminishing with time beyond-separation).

6.7.1 Adjusting payments for risk factors and beyond-separation effects

In funding, like performance measurement, disutility events beyond-care can be added as additional factors, before considering payment conditional on these rates. However, unlike performance measurement, relative costs beyond-care cannot be added, but rather, need to be adjusted for. This is most easily achieved by including costs beyond separation directly as disutility events with a WTP of one dollar, or equivalently

converting them to equivalent disutility event rates, at the scheduled WTP, for such disutility events.

In adjusting for risk factors, at an individual clinical activity (DRG) level, problems of case-mix differences in mix of clinical activities are avoided. However, providers can still have differences in within-activity case-mix, facing populations with different expected costs and disutility event rates, given differences in patient risk factors. Adjustment for factors outside of the control of hospitals, such as patient age, sex, medical history and diagnosis, are required to allow for differences in expected disutility event rates and costs prior to calculation of payment conditional on disutility event rates. This can be undertaken, as described in section 5.7.1, using regression methods to estimate expected rates, and then standardizing relative to means. It should be noted that adjustment of payments for within-DRG patient differences is already required in considering expected cost differences per admission in case-mix funding. The proposed method, therefore, imposes no additional burden, with respect to adjusting expected costs for patient factors at a DRG level, than current methods.

As in performance measurement, to retain covariance structure between costs and disutility events rates (quality of care), the same data should, however, be used in calculating expected costs as expected disutility event rates. However, unlike performance measurement adjustment of payments for differences in expected costs per patient and disutility event rates, given covariates need to be undertaken by standardization. This requires using differences in individual expected hospital costs relative to industry expected costs.

6.8 Summary

A question posed in chapter 1 was whether appropriate incentives for quality of care could be systematically created in funding hospitals, using current evidence and within current budgets. In this chapter, funding conditional on relative disutility event rates, where correspondence conditions are satisfied, has been shown to provide a systematic framework to trade-off the cost and value of quality of care, consistent with maximising net benefit per admission. A sequential two stage funding mechanism has been illustrated that allows budgetary constraints, at current case-mix payments per

admission, to be retained in moving from a case-mix payment system, ignoring quality of care, towards net benefit maximisation.

Activity-based funding systems for hospitals, such as DRG based case-mix payments, in ignoring quality of care, were demonstrated to create economic incentives for reducing quality of care until the expected marginal cost of reducing quality of care within admission is 0. These economic incentives can be opposed to varying degrees by: attempts at regulation of hospitals, such as health technology assessment, accreditation and clinical performance monitoring and; health objectives of providers and other social institutions, to the extent they can influence resource decisions. However, in failing to hold hospitals economically accountable for quality or effects of care, case-mix funding mechanisms also allow technical inefficiency to be hidden behind lower quality of care, where quality of care at an industry level is above that of cost minimisation. That is, where arguments for a non-zero value for health objectives prevail, in spite of cost-minimising economic incentives from case-mix funding. At a health system level, the detrimental effects of lower quality on health outcomes create the need for increased care post-separation. These effects reinforce the perversity of encouraging cost-minimising quality of care within admission.

Recent initiatives to condition a proportion of payments on reaching threshold levels of quality of care, while improving on a zero value of quality of care, have no economic basis for determining the proportion of funding. They can also only potentially provide appropriate incentives in trade-offs between costs and quality of care, at localised threshold levels.

In identifying continuous net benefit maximisng incentives for quality of care, the ordinal correspondence between maximising net benefit per admisison and minimising cost plus disutility events per admission, valued at the decision makers value of willingnes to pay identified in chapter 5, has been demonstrated to also be cardinal. This cardinal correspondence allows funding conditional on differences in disutility event rates, consistent with net benefit maximisation, where correspondence conditions are satisfied.

To allow a transition from case-mix funding to a funding mechanism conditional on quality of care a two-stage payment mechanism has been proposed with:

1. an initial scheduled payment relative to best practice, estimated with frontier methods, conditional on the scheduled value for quality and;
2. a residual payment pro-rata per admission to act as a buffer.

Hospitals with low quality of care, and particularly those that have hidden technical inefficiency behind low quality, face funding shortages relative to current payments, unless they increase quality. In making hospitals accountable for quality, as well as cost of care, economic incentives are provided to convert such inefficiencies into increased quality of care.

To allow for hospitals to adapt to economic accountability for effects of quality within current case-mix funding budget, a sequential 2-stage funding mechanism was illustrated. An initial scheduled value of quality set at the estimated industry value of quality (shadow price) ensures current budgetary constraints per admission are met, unless quality improvement is greater than technical inefficiency. The value of quality is then sequentially moved towards that of net benefit maximisation. As provider behaviour changes in response to incentives for quality of care, shadow price for quality across the industry should broadly change to reflect the scheduled value.

In transition towards a system of funding based on evidence-based medicine, the sequential funding schedule has been illustrated to allow a funding mechanism which:

1. provides more appropriate incentives at each stage;
2. is conducive to planning and management of administrators and;
3. maintains budget constraints.

Compared with a one-step mechanism, this should enable a manageable culture change in moving towards a schedule which reflects and creates incentives for net benefit maximisation. Consideration of losses in net benefit from inappropriate incentives, relative to the problems of adjusting to accountability for effects of quality of care have been suggested in identifying an optimum rate of increasing from a 0 value to that consistent with maximising net benefit.

Estimating potential gains in net benefit attributable to overcoming technical, allocative and scale inefficiency were illustrated by comparing the position of isocost curves

relative to best practice in the cost disutility plane. Data linkage to adjust for effects beyond-care, patient level data to allow adjustment for patient risk factors within-DRG and effects of sample variation were, however, stressed as potential limitations in these illustrations.

To satisfy correspondence conditions, and systematically allow incentives for net benefit maximising quality of care, requires meeting existing policy challenges for robust risk factor adjustment with patient level data, and data linkage to effects beyond-care. The cost and benefits of meeting these policy challenges, in allowing robust application of the correspondence framework to performance measurement and funding at a clinical activity (DRG) level, are explored in chapter 7. Methods to allow for potential effects of sampling variation on the position of the frontier are considered in future directions in chapter 8.

Chapter 7: Policy Implications

7.1 Overview

Funding hospital inpatient services based on case-mix adjusted admissions and measuring hospital performance with activity-based measures, such as cost per case-mix adjusted separation, implicitly includes the costs, but ignores the effects, of quality of care. This failure to account for effects of care creates economic incentives for cost minimising quality of care, opposing the use of evidence-based medicine in practice. The correspondence theorem provides policy makers with a method for flexibly including disutility events (effects of quality of care) in performance measurement and funding, consistent with an underlying objective function of maximising net benefit per admission, where correspondence assumptions are satisfied.

To satisfy correspondence conditions, and overcome incentives for cost-shifting and cream-skimming, existing policy challenges remain in both data linkage to effects beyond-care (to prevent incentives for cost and event-shifting) and adjusting for patient risk factors at admission (to prevent incentives for cream-skimming). Use of decision analytic methods (Weinstein and Fineberg, 1980; Pettiti, 1994; Hunink et al., 2001), at a clinical activity level of analysis, allows these perverse incentives to be overcome within an evidence-based framework. Disutility events within, and beyond, care can be flexibly identified by DRG and differences between hospitals in within-DRG patient risk factors can be adjusted for. The linear nature of correspondence allows easy synthesis of multiple disutility events, including those beyond point of separation with data linkage. Perceived utility bearing effects of care can be reframed and included as disutility events where they can be expressed as utility, cardinal measures of functioning or utility bearing events, as described in section 5.7.2.1. Where combinations of disutility events are present in patient populations, they can be valued separately using decision analytic methods to allow for interaction in valuation, as described in section 5.7.2.6.

As data linkage is already required in attempting to allow for cost-shifting under case-mix funding, resources to additionally allow linking of effects beyond-care can be

considered at the margin.[1] However, the value of the pay-off from allowing performance measurement and funding under correspondence conditions, are considerable. Applying the proposed performance measurement and funding framework, under correspondence conditions, current incentives for cream-skimming, reduction in quality of care and cost-shifting can be replaced by the maximisation of net benefit per admission in treating patient populations within a health care system. While valuing relative effects of care creates incentives for allocative efficiency in net benefit maximisation, the associated accountability removes the ability to hide technical inefficiency behind lower quality of care. The sequential two-stage funding mechanism, developed in chapter six, allows accountability for quality to be introduced in a managed way with more appropriate incentives provided, while maintaining funding at case-mix levels per admission. Policy-makers, as principals, can create economic incentives for hospitals, as agents, consistent with maximising net benefit in practice.

In considering effects on the internal relationship and behaviour within hospitals, with funding conditional on quality of care, administrators can no longer act simply as accountants in minimising costs within-admission. Administrators face a trade-off between the costs and the policy maker's value of quality (effects) of care, while clinicians become accountable for the effects of their quality of care, relative to best practice. The objective of maximising net benefit under this funding mechanism, in valuing and accounting for effects of quality of care, creates a rational evidence-based framework to address current tension between clinicians attempting to maximise health and hospital administrators minimising costs.

7.2 A policy framework for appropriate quality of care

In health technology assessment, decisions are increasingly informed by evidence of the relative effect and costs of technologies. Economic evaluation of new technologies, allowing for evidence of relative clinical effectiveness and cost (resource) implications, has, during the past decade, been required in Australia in

[1] In comparison with the lack of co-ordination of current measurement and monitoring of effects in accreditation, peer review, clinical audit and hospital performance measures at different levels of aggregation, the systematic nature of the proposed framework at a clinical activity level removes potential duplication. Resources devoted to monitoring could therefore even be potentially reduced.

evaluating pharmaceutical products by the Pharmaceutical Benefits Advisory Committee (PBAC,1995) and more recently in evaluating medical technologies by the Medical Services Advisory Committee (MSAC). This approach of 'evidence-based medicine' has also been adopted in undertaking health technology assessment in the UK by the National Institute of Clinical Excellence (NICE) and in Canadian provinces such as Ontario. Decisions made under this approach, trading off incremental costs and effects relative to a decision making threshold for the value of effects of care, reflect a policy objective of maximising net benefit (Stinnett and Mullahy, 1998).

However, while policy makers increasingly recognise, and are informed of, evidence of trade-offs between costs and effects of care in relative performance of new technologies, case-mix funding and performance measurement in hospitals continue to focus on costs of care alone. For any given clinical activity, technology is assumed to be uniformly applied across providers with no differences in effects of care across providers in the choice, or use, of technology. Reducing cost per admission is measured as improved performance, regardless of effects on quality of care. As described in chapter 2, and illustrated in chapters 5 and 6, this creates economic incentives for reduction in quality of care below a net benefit maximising level (allocative inefficiency). The lack of accountability for quality of care also creates the scope to hide technical inefficiency behind lower quality of care, and to shift costs beyond point of separation.

For hospital performance measures and funding mechanisms to reflect an objective function that supports evidence-based medicine in practice, the value of effects from quality of care needs to be traded off against the cost of quality of care.

A DRG (clinical activity) level of analysis allows health effects to be flexibly identified and attributed with disutility event rates such as mortality, morbidity, readmission and iatrogenic events, appropriate to each clinical activity. A clinical activity level, therefore, provides the ability to flexibly and comprehensively include the value, as well as costs, of quality of care, in overcoming the 'performance-efficiency paradox' (Perkuninnen et al., 1991). As described in chapter 3,

undertaking performance measurement at a clinical activity level also avoids the Fox (1999) aggregation paradox, and identifies inefficiency hidden by aggregation.

Output specifications of disutility events are, however, unable to provide an economic efficiency measure with an appropriate trade-off between cost (resource use) and value (effects) of quality of care, as illustrated in chapter 4. The hyperbolic approach of Färe Grosskopf, Parsuka and Lovell (1989) allows performance measured in equi-proportionally expanding strongly disposable desirable outputs (such as electricity), and contracting weakly disposable undesirable outputs (such as pollution). However, this method does not translate to performance where disutility event rates represent quality of care and admissions *per se* are not desirable. In figure 7.1(a), specifying admissions as desirable outputs and disutility events as undesirable outputs, technical efficiency is estimated with the method of Färe Grosskopf, Parsuka and Lovell (1989) in radially contracting disutility event and expanding admissions. However, as argued in section 4.7, technical efficiency estimated relative to regions of the frontier such as CD, becomes meaningless as a performance measurement where disutility event rates standardised at a clinical activity level reflect quality of care, rather than differences in patient populations or other external influences. This is particularly problematic as output orientated economic efficiency (e.g. revenue efficiency) is not estimable in the absence of prices for admissions or disutility events.

Congestion inefficiency is able to be interpreted in the case of pollution, as a by-product from production such as electricity generation as the additional desirable output able to be produced without constraints on undesirable outputs. However, in hospitals, congestion inefficiency does not have a valid interpretation where admissions *per se* are not valued, without consideration of associated quality of care (disutility event rates). Congestion inefficiency as the increase in admissions without constraints on disutility event rates directly opposes the derived nature of demand for admissions in improving health.

Figure 7.1 (a) Technical and congestion efficiency with undesirable events as outputs

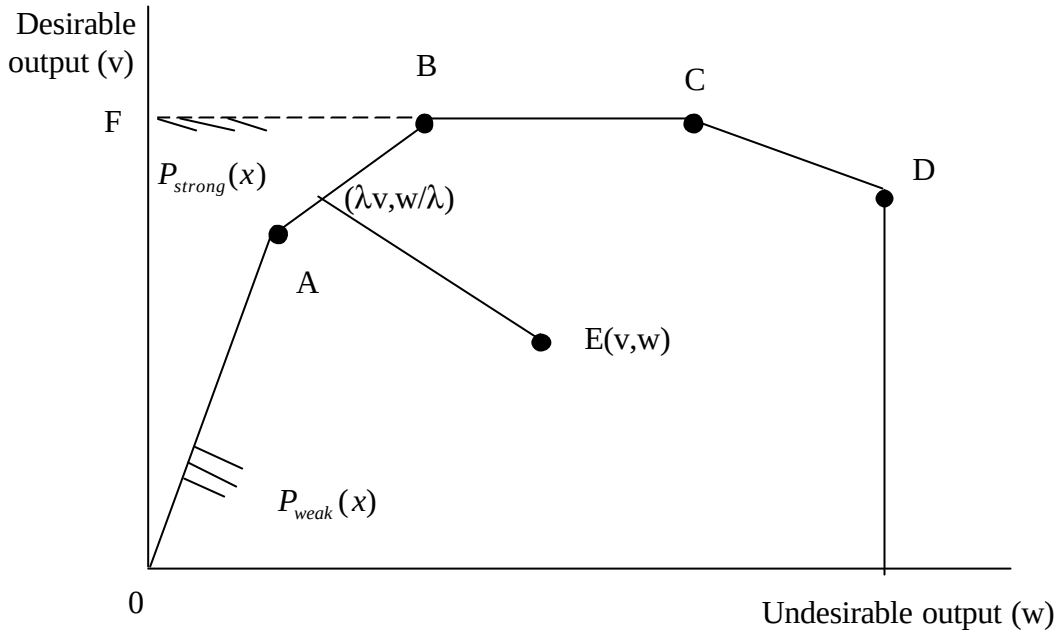
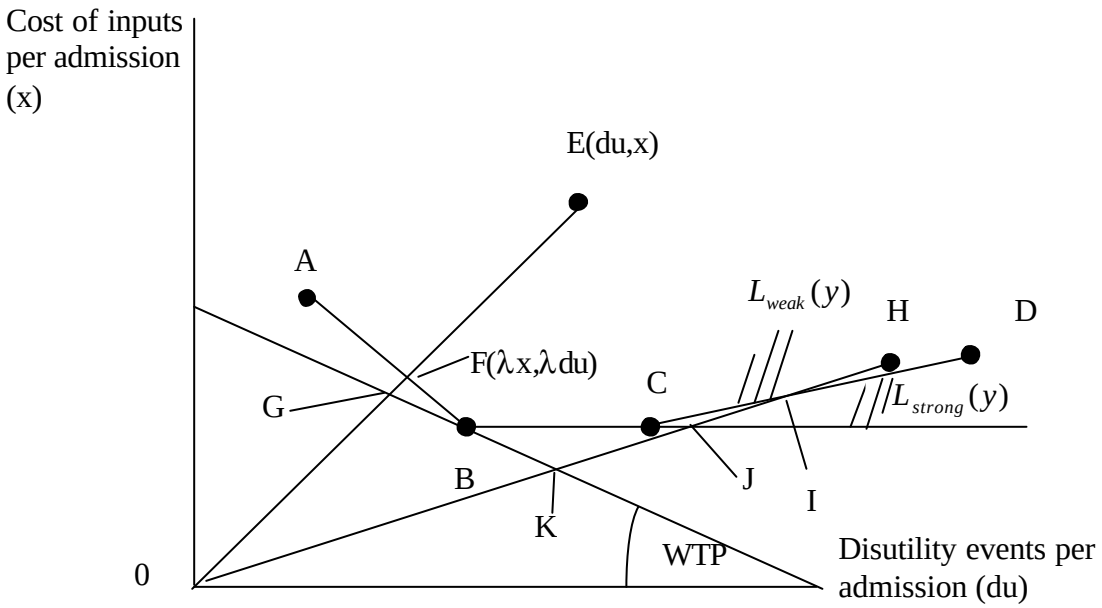


Figure 7.1 (b) Technical, economic, allocative and congestion efficiency with undesirable events as an input under constant returns to scale



In comparison, in specifying disutility events as an input in figure 7.1(b), for a hospital at E, technical efficiency is estimable as 0F/0E, economic efficiency (consistent with net benefit maximisation under correspondence conditions) as 0G/OE and allocative efficiency as 0G/OF. For a hospital at H (above cost minimising disutility event rate) congestion efficiency can be estimated as OJ/OI.

The hyperbolic method does not allow meaningful estimation of:

(1) economic or allocative efficiency;

(2) shadow prices in the absence of prices for marketable desirable outputs (admissions in the model) and;

(3) congestion efficiency as the residual of technical efficiency under weak and strong disposability of disutility events.

Technical efficiency measured relative to regions of the frontier, where, for given resources admissions fall and disutility events increase, is also theoretically flawed as a performance measure, unless disutility events are assumed to be determined outside hospitals' control. In comparison, application of the correspondence theorem allows meaningful and intuitive estimation of economic, technical, allocative and congestion efficiency.

As illustrated in chapter 5, applying the correspondence theorem to performance measurement at a clinical activity level allows:

(1) Identification of potential best practice peers conditional on quality of care;
(2) estimation of industry shadow price for effects of care;
(3) estimation of economic efficiency consistent with maximising net benefit;
(4) estimation of technical efficiency reflecting dual objectives of cost minimisation and health maximisation and;
(5) allocative efficiency as a residual of economic and technical efficiency reflecting appropriateness of the implicit valuation of quality of care relative to net benefit maximising best practice.

Under correspondence conditions, these relative performance measures are consistent with maximising net benefit per admission, providing an appropriate trade-off between the value of health effects and cost of quality of care.

In satisfying correspondence conditions, a clinical activity (DRG) level allows decision analytic methods to be used, as in health technology assessment, to flexibly identify health effects and costs within, and beyond, separation, as described in section 5.7.2. Comprehensive coverage is also facilitated by the ability of the correspondence theorem's linear form to handle multiple events, including those beyond-separation and to reframe utility bearing aspects of care as disutility event 'rates'. Application of the correspondence theorem to performance measurement provides both a method, and a framework, for policy makers to identify best practice and relative performance, allowing for the effects, as well as costs, of quality of care.

Performance measurement can provide appropriate incentives in benchmarking and peer identification. To provide appropriate economic incentives for quality of care in practice requires hospital funding mechanisms to reflect an objective of net benefit maximisation. In allowing a net benefit maximising funding mechanism, the ordinal correspondence, identified in chapter 5 for relative performance measurement, was shown to also be cardinal, in chapter 6.

Differences in net benefit correspond to differences in minimising cost plus disutility events valued at the decision maker's value of willingness to pay, as in net benefit. Under correspondence conditions, this cardinal correspondence allows funding consistent with maximising net benefit by conditioning payments on differences in standardised disutility event rates.

7.2.1 Policy challenges

For policy makers, the correspondence theorem provides an evidence-based framework to systematically replace perverse economic incentives for cost minimising quality of care, with accountability for effects and cost of quality of care under an appropriate trade-off. This framework, while allowing comprehensive and flexible identification of what is required at a DRG level, and providing a linear and flexible method for synthesis, is not, however, a panacea by itself. Satisfying correspondence assumptions focuses existing policy agendas on risk adjustment, to avoid cream-skimming, and data linkage to avoid cost and effect shifting.

To satisfy the common comparator assumption, and overcome incentives for cream-skimming, challenges remain for policy makers in adjusting costs and disutility event rates for patient risk factors. To curtail incentives for cost and event shifting, and satisfy the coverage of effect assumption, requires linkage to costs and effects beyond-separation or measuring verifiable effects of care at point of separation.[2]

However, adjustments for patient risk factors, and the inclusion of effects beyond hospital admission, are required to be systematically addressed by any performance measure and funding mechanism to avoid incentives for cream skimming and cost and effect shifting. Similarly, partial attempts to avoid these incentives, such as peer review or monitoring of clinical effects, accreditation, or any regulation of performance, need to allow for risk factors and effects beyond-separation.

7.2.2 Investing in data linkage and risk modelling

Policy makers generally recognise the importance of adjusting for risk factors at admission, and effects beyond-separation, but lack a theoretical framework to appropriately synthesise these effects, in considering performance measurement and funding. In Australia, various initiatives in data linkage have been, and continue to be, undertaken in modelling effects of hospital admissions beyond point of separation and interactions between hospitals and other health care providers. These initiatives include:

(1) linkage of hospitalisation to residential aged care by the Australian Institute of Health and Welfare;

(2) linkage of Medicare data, for patient level prescription drug use under the Pharmaceutical Benefits Scheme (PBS) and general practitioner and other Medicare services under the Medical Benefits Schedule (MBS), by the Health Insurance Commission (HIC);

(3) Western Australian data linkage (using probabilistic modelling) of individual patient hospitalisation morbidity (re-admission by principal diagnosis) death, cancer and mental health registry data back to 1971 (Holman, Bass, Rouse and Hobbs, 1999; Kelman, Bass and Holman, 2002), combined with linkage

[2] Effects measured over time may also be appropriate in some hospital activities, particularly for palliative or non acute care.

of residential aged care and the HIC MBS and PBS data in more recent years;

(4) Commonwealth Government co-ordinated care trials;

(5) The health-connect electronic health record used clinically (matching with unique identifiers) and;

(6) The current Australian Health Ministers Advisory Council commitment to extend the data linkage approach in Western Australia nationally.

Despite these data linkage initiatives, quality or effects of care continue to be excluded from hospital inpatient performance measures and funding mechanisms. Perverse incentives for quality of care, therefore, remain in the absence of a framework, and method, to systematically include the effects and costs of quality of care consistent with an appropriate objective function.

The correspondence theorem, applied at a clinical activity level in performance measurement and funding, provides such a systematic evidence-based framework. Where correspondence conditions are satisfied, appropriate economic incentives for quality of care are created, with an underlying objective of net benefit maximisation, and perverse incentives for cost shifting and cream skimming are avoided.

Incorporating data linkage to costs alone, under case-mix funding, is not enough to provide appropriate incentives for quality of care, and could even result in incentives for practices leading to worse health outcomes. For example, perverse economic incentives to increase mortality rates may be created, to prevent cost-shifting from sick patients. However, as disutility events such as deaths are explicitly allowed for in the proposed framework applying the correspondence theorem, these perverse incentives would be removed.

Application of the correspondence theorem comes at a cost of identifying, and obtaining data, to meet requirements for measuring effects within-care and data linkage to effects beyond-care at a clinical level. In Australia, the current linkage of patient data in Western Australia, and initiatives to extend this data linkage approach to other states, suggests that, in satisfying correspondence conditions, any investment would be at the margin.

The systematic evidence-based focus in satisfy correspondence conditions, using decision analytic methods, at a clinical activity level, suggest data linkages to satisfy coverage of effects, once set up, would not require replication. In comparison, the partial and disjointed nature of current coverage, with separate links to costs and clinical effects (such as accreditation, peer review, clinical audit), at different levels of aggregation, replicates effort, while still providing scope for finessing by providers in quality of care aspects not included in funding. Such partial coverage, without an underlying unifying framework, characteristically creates a process of continual reform in attempting to overcome perverse incentives from partial measures. Replication of effort with partial measures can, therefore, be expected over time, as well as at different levels and scopes of coverage, at points in time. Applying the proposed correspondence theorem framework using decision analytic methods at a clinical activity level is both systematic and comprehensive. Avoiding current duplication across levels and types of monitoring and processes of continual reform may, therefore, over time, even reduce resources devoted to risk factor adjustment and monitoring of effects beyond point of separation.

Even if there were a net investment of marginal resources in monitoring to allow effects, as well as costs, to be linked, and risk factors modelled, the relative advantages of using the correspondence theorem framework are considerable. Where correspondence conditions are satisfied, the framework allows the ability to create incentives for appropriate quality of care, while avoiding cream-skimming and cost-shifting in performance measurement and funding.

In the principal agent terms of Goddard, Mannion and Smith (2000), the correspondence theorem, applied to performance measurement and funding at a DRG level using decision analytic methods, provides an appropriate and consistent framework for measurement, attribution and reward. Where correspondence assumptions are satisfied, performance measurement and funding mechanisms applying the correspondence theorem create economic incentives for the agent (hospitals as providers) to maximise net benefit. This reflects the economic objective of the principal (government as purchaser of health services). The objective function of net benefit maximisation, in treating a patient population, represents an appropriate trade-off between objectives of health maximisation and cost minimisation for

hospitals. Following Harris (1977), the question, of what effect moving from an objective of cost minimisation to net benefit maximisation may have on the internal hospital relationship between clinicians and administrators, should also be considered.

7.3 Effects on internal hospital relationships in moving from cost minimisation to net benefit maximisation

Rather than a single agent, the hospital is suggested by Harris (1977) to be organized as two firms, with administrators attempting to minimise costs per admission (under case-mix or other activity based funding mechanisms) and clinicians attempting to maximise health outcomes of their patients treated. Harris identifies the central feature of hospital production as a separation of internal supply and demand functions. Administrators, while not making clinical decisions themselves, plan capacity for inputs and respond to internal demands from individual clinicians. They arrange and co-ordinate supply of various ancillary services such as x-rays, pharmacy, operating rooms or theatre, blood bank and beds, given demands made for these services by clinicians, on behalf of patients treated.

Harris (1977:469) suggests the reason for this separation is because hospital care often involves a complicated sequence of adaptive responses in the face of uncertainty, given patients can have numerous idiosyncrasies or risk factors, and failure to take the necessary actions at precise times can have disastrous consequences. Uncertainty in diagnosis of patients, and hence their resource requirements, can lead to highly inelastic demand, when revealed. Harris consequently suggests that:

> "Diagnosis of treatment and illness …. requires an organization which can adapt rapidly to changing circumstances and new information… hospitals have developed a specialized system of very short run internal resource allocation to handle this co-ordination problem." Harris (1977:469)

Administrators with an objective of cost minimisation, ascribe a high weight to losses from holding excess capacity of inputs, while clinicians, with health maximisation as an objective, attach a high weight to potential health losses from lack of capacity, given uncertain, and highly inelastic, demand for inputs in treating any given patient.

Characteristically, a run down of excess capacity (by hospital administrators with a high weight ascribed to holding excess capacity) results in clinicians engaging in a 'mad scramble' to ensure resources, to defend quality of care under highly inelastic demand, in treating their patients. Where supply of resources in treating patients becomes scarce doctors can:

> "deem all sorts of demand as necessary for their patients." Harris (1977:477)

Consequently, Harris suggests that:

> "Hospitals with apparent capacity excess or cost overruns may actually be in a deceptively stable equilibrium." Harris (1977:480)

Where funding ignores quality of care and is based on activity alone (as in case-mix funding), excess capacity is measured as technical inefficiency. Incentives are created for administrators to reduce excess capacity towards 0. However, as argued by Harris (1977), excess capacity of inputs is required to allow quality of care received by patients under uncertainty, given inelastic demand and supply, and particularly in the very short run decision making context of hospital care. Excess capacity may, therefore, be seen as required for quality of care but also to prevent inefficient hoarding practices by clinicians in response to breakdowns in the allocation mechanism:

> " … when the medical staff's internal demands exceed the short run capacity supplied by the administration." Harris (1977:467)

In the proposed funding mechanism, applying the correspondence theorem, there is an explicit trade-off between the value of the effects (disutility event rates), and costs, of quality. Under correspondence conditions, economic incentives are created for a net benefit maximising level of excess capacity, given the implicit expected value and cost of excess capacity.

Under case-mix funding, 'clinical neutrality' can only be maintained by the organization of the hospital itself, in separating clinicians, and their decision making and demands, from administrators, and their decisions with respect to supply of inputs, in meeting clinical demands. As Harris characterizes, this internal splitting of the hospital attempts to allow doctors to act as agents for patients alone:

> ".. it should be understood the organization is set up to protect the doctor from behaving as economic man." (Harris,1977:468)

However, as Harris later suggests, this splitting of the organisation, results in many inefficiencies:

> "The failure to recognize that doctors and hospitals are linked by a strong bond of joint production is the basis of many of the hospitals inefficiencies."
> (Harris, 1977:475)

These inefficiencies include those that relate to a lack of knowledge in trade-offs between cost and quality of care in both sides of the split in internal organisation, as well as those related to breakdown in allocation mechanisms for inputs and hoarding practices of clinicians.

Inefficiencies related to the lack of recognition of joint production in decisions making under organizational separation, are the basis for the ultimate conclusion Harris (1975:467) reaches that: "ultimately a rational public policy towards hospitals requires a change in the internal organisation of the hospital itself."

In a related conclusion, Harris also suggests that the primary objective of health maximisation, and quality, or doctor part of the 'hospital firm', need to be included alongside the objective of cost minimisation where currently:

> "policy towards hospitals is almost exclusively directed at the supply side of the organisation….this policy is doomed to failure".
> (Harris,1977:467)

Including heath effects consistent with the economic objective of net benefit maximisation, under the correspondence theorem, means this primary policy concern of Harris (1977) for inclusion of health maximisation in the objective function, can be directly addressed, in performance measurement and funding.

Use of the proposed performance measures and funding mechanisms, described in chapters 5 and 6 respectively, allow policy makers to create a structural framework for a shift in hospital organisation. Clinicians and administrators are provided with

guidance in valuing quality and a common objective, which trades off the cost and value of quality of care. This structural framework allows hospitals to be organized such that clinicians and administrators can have a dialogue in relation to the common goal of net benefit maximising quality of care under funding constraints.

The funding mechanism acts as a structure for the relationship between administrators and clinical decision makers, in influencing both the objective, and the constraints, of the negotiation framework. While Harris (1977) characterises administrators as cost minimising, the objective function that underlies the funding mechanism largely determines the primary argument in the administrator's objective function. Under case-mix funding, the administrator's underlying objective is implicitly cost minimisation per admission. Conditioning payments on disutility event rates, at the decision maker's value of WTP, under correspondence conditions, the administrator's underlying objective becomes net benefit maximisation.

The objective function implicit in the funding mechanism also affects clinicians' negotiation position in demanding resources, to the extent it reflects, and is informed by, clinical effects. Under case-mix funding, clinician demands for inputs are implicitly valued up to a quality of care minimising cost per admission, but are only constrained at a hospital level and relative to average expected industry costs.[3] Under the proposed funding mechanism, where correspondence conditions are satisfied, clinician demands for inputs are valued up to a quality where net benefit is maximised and relative to observed best practice for net benefit maximisation in that clinical activity.

Under case-mix funding, any negotiation between clinicians and administrators occurs with an underlying objective of cost minimisation, providing economic incentives for cost minimising quality of care for each clinical activity (DRG). However, this will not necessarily lead to costs minimising resource use per admission at a clinical level, given inefficiencies created by the internal split, between clinicians and administrators, that this system requires. Additionally, at a hospital or industry level, costs minimising resource use per admission may not result, as outlined in chapter 6,

[3] Value is also potentially justified in meeting minimum accreditation standards.

given the scope for hiding technical inefficiency behind quality of care, where payments reflect expected average cost per admission across industry and do not hold hospitals accountable for quality of care.

The proposed funding mechanism creates a more appropriate trade-off between the value of quality of care (currently implicitly valued at 0 by case-mix funding) and cost of quality of care. As payments are relative to best practice, this creates economic accountability for quality of care. Under correspondence conditions, this funding mechanism creates a structure for the internal hospital relationships between manager and physician to support quality of care that maximises net-benefit in practice. The mechanism supports evidence-based medicine in terms of both how, as well as which, strategies or technologies are adopted, in treating patients for any given clinical activity (DRG).

Administrators, under the proposed framework, can no longer simply be cost minimising accountants for any given clinical activity (DRG). Rather, they are required to consider opportunity costs that resource allocation decisions represent at the margin, in terms of expected effects (represented by disutility event rates) from quality of care. Health care providers (clinicians), while having their effects of care valued, become economically accountable for these effects of care relative to net benefit maximising best practice. Economic incentives and requirements are therefore created to adopt quality of care reflecting net benefit maximising, evidence-based medicine in practice.

Moving from a funding system with an objective of cost minimisation to one where the objective is net benefit maximisation, can only aid a more appropriate economic basis to the internal negotiation of quality of care, and the relationship between administrators and clinicians generally. The conditioning of payments on performance, in meeting health objectives, relative to best practice, provides an environment for internal negotiation in resource allocation, with active incentives in each hospital for a process akin to program budgeting and marginal analysis (Mooney, Russell and Weir, 1986; Mooney, Gerrard Donaldson and Farrar, 1992; Mooney, 1994a:27-32).

Similar to program budgeting, the framework encourages clinicians, and administrators, to review how resources are being used, against effects of care for specific objectives and at a disaggregated output orientated program (DRG) level. As in marginal analysis, for given resources, the output orientation in conditioning of payments encourages shifting of resources at the margin, within, and between, programs, if there is an increase in value of benefit in treating a patient population, and hence net benefit. For each clinical activity, the objectives, and their degree of importance, are implicitly set by policy makers in the effects of care included and relative value scheduled payments place on these effects of care. As in program budgeting, the identification of objectives should be undertaken in consultation with health care providers, and patients, to allow informed values and ownership of the process.

7.4 Policy implications for payment systems

Ideally, hospital funding mechanisms would allow policy makers to provide net benefit maximising incentives for quality of care with verifiable measures of effects of care, while maintaining budgetary control and equity of access to appropriate care. In the following sections, comparison of case-mix funding to the proposed funding mechanism, in meeting each of these objectives, are considered.

7.4.1 Funding to allow appropriate incentives and budgetary control

Payment systems based on hospital activity, *per se*, provide active economic incentives to hospitals, as agents, to engage in practices of cream-skimming, cost-shifting and reducing quality of care in minimising cost per admission. While case-mix adjustment can remove some scope for cream-skimming, adjustment undertaken at an aggregate level focuses on between-activity average cost differences, rather than within-activity patient risk factors.

Economic incentives for cost minimising quality of care, under case-mix funding, may be opposed for any given clinical activity, and any hospital, to a lesser or greater degree, by health arguments in objective functions for clinicians and administrators and their internal relationship within the hospital in decision making. The degree of opposition to cost minimisation, for any given clinical activity within any hospital,

depends on discretionary influences of objectives of clinicians and administrators and the influence of constraints with attempts at regulation of clinical activity.[4] Average costs for each DRG, on which case-mix weights are based, reflect variable quality of care above that of cost-minimisation in practice.[5] Consequently, as discussed in section 6.2, case-mix payments, based on historical average costs across sampled hospitals, allow technical inefficiency to be hidden behind lower quality of care by individual hospitals, to the extent that, in practice, industry quality of care lies above levels minimising cost per admission.

In considering health system effects beyond-separation, case-mix payments, while actively creating incentives for care at a quality below net benefit maximisation, do not provide budgetary control in treating a patient population over time across a health system. Budgetary control in treating patient populations, at a health care system level, may be provided by methods of capitation-based funding in geographic regions, often at regional levels and in conjunction with needs-based criteria, as discussed in section 7.4.3. [6]

Case-mix funding, in summary, creates economic incentives for cost minimising quality of care and cost-shifting, while still providing scope for technical inefficiency to be hidden behind lower quality of care and not accounting for patient risk factors within-DRG. In comparison, funding conditional on quality of care applying the correspondence theorem, allows incentives for net benefit maximising quality of care, while avoiding cost-shifting, and cream-skimming, incentives under correspondence conditions.

In applying the correspondence theorem, a sequential two-stage funding mechanism was identified and illustrated in chapter 6, allowing policy makers to create a smooth process of change, in moving from an objective of cost minimisation toward net

[4] As section 5.2.2 and 6.2 described, system processes of health technology assessment, accreditation, peer review and monitoring of clinical performance, influence of other social institutions and the Hippocratic oath and other professional arguments for health in providers objective function may mitigate against economic incentives for cost minimisation to differing degrees across hospitals.
[5] In Australia, Australian National Diagnostic Related Group (AN-DRG) weights are estimated each year by the National Hospital Cost Data Collection (CDHA (2000)) across a sample of public and private hospitals by type described in section 1.4.
[6] Rice and Smith (2001) provide an international survey and discussion of such capitation approaches.

benefit maximisation. This proposed sequential two-part funding mechanism allows increasingly appropriate incentives for quality of care, while maintaining budgetary control per admission. The key components in transition from case-mix funding are:

(1) First-stage payments conditional on differences in disutility event rates relative to best practice at an implicit scheduled value for quality (avoiding disutility events);

(2) Setting the initial schedule value at, or above, current industry shadow price for disutility event rates (quality of care) and sequentially moving towards a net benefit maximising value[7], and;

(3) To smooth over transition from case-mix funding a second stage "buffer" payment per admission (not conditional on quality of care) is made representing the residual of case-mix funding.[8]

Gradually increasing the value of avoiding disutility events towards that of net benefit maximisation, allows hospitals, administrators and providers (clinicians) to adapt to accountability for quality of care, and the increasingly appropriate incentives this provides. The mechanism allows incentives for quality of care to change in a planned and managed way, while maintaining budgetary control in retaining average industry levels of case-mix funding per admission.

Appropriately, the greatest economic imperatives to improve quality are provided to hospitals that have hidden technical inefficiency behind low quality of care, who otherwise face budget shortfalls. In conditioning payments, relative to best practice, the funding mechanism both values, and creates, accountability for quality of care. Valuing quality addresses allocative inefficiency in net benefit and accountability addresses technical inefficiency hidden behind low quality of care. Under correspondence conditions, this mechanism should improve net benefit per admission at an industry level, but may also provide better budgetary control in the treatment of given patient populations, across the health system over time. Higher quality of hospital care, to the extent that it leads to better health effects beyond-separation,

[7] The industry shadow price for quality is predicted to be lower than the decision maker's net benefit maximising value, given economic incentives for a 0 value under case-mix funding.

[8] The scheduled value of avoiding disutility events can be increased until the net benefit maximising value is reached or the buffer payment reaches 0. Initially using the current shadow price of quality in

reduces the expected need for health care post-separation, with the potential exception of where improved quality reduces mortality.

To manage the rate of change in funding schedules, and ensure performance is responding appropriately, policy makers can monitor and verify performance.

7.4.2 Verifying and monitoring performance over time

Policy makers can monitor performance, at a clinical activity level, in response to incentives created by application of the correspondence theorem in funding, using the performance measurement framework described and illustrated in chapter 5. Using this framework in considering effects over time, policy makers can be informed over time of:

(1) Relative valuation of quality of care (shadow price) of industry behaviour where industry economic efficiency is maximised, as illustrated in section 5.6.4;

(2) Change in net benefit conditional on value of avoiding disutility events, as demonstrated in section 6.6, and;

(3) Degree of convergence of behaviour relative to best practice, summarized by indicators such as average economic and technical inefficiency across providers, variance, and identification of outliers.

Comprehensiveness, in measuring effects of quality of care, is important to prevent a focus on satisfaction of measured aspects at the expense of other quality of care aspects. Identifying quality indicators at a DRG level, using decision analytic methods, allows a comprehensive and integrated approach to coverage. However, the verifiability of measures also has to be considered.

To prevent gaming in reported measures of effect, less verifiable aspects of quality may require feedback loops, from processes such as random sample clinical audit and peer review, as outlined in section 5.7.2.1. Efficient design of such monitoring requires consideration of the costs, relative to the benefits, of monitoring, and consequently the optimum size of random samples in audit processes. Optimum

---

industry behaviour as an initial schedule value of quality ensures the average level of payment per

verification approaches and sampling size should be flexible, in reflecting potential differences across activities (DRGs), including the degree of verifiability of quality measures, volume of patients, and the nature of clinical culture.

Applying the correspondence theorem, as illustrated in chapter 5, allows an explicit consideration of trading off the cost and value of quality and activity in relative performance, at a point in time. However, in general, comparison of performance over time with cost data is problematic, and should be avoided if possible. In assessing changes in relative efficiency (technical efficiency) and movement of the frontier (technical change) over time, physical inputs, rather than costs, would ideally be used, as described in section 5.5.4. If only costs are available in longitudinal analysis, they need to be deflated to a base year, to allow for price changes over time. As outlined in appendix 4.2, to estimate technical efficiency with cost data using radial contraction property under DEA, requires the assumption of all firms facing the same factor prices at each time period.

If data on input factors were available, then productivity change over time could be calculated using Malmquist index methods, as outlined in appendix 7.1, with admissions as an output and physical factors of production as well as disutility events (representing quality of care) as inputs.

In considering use of Malmquist index methods, allowing for quality over time, Färe, Grosskopf and Roos (1995) developed a method for the incorporation of positive non-traded quality attributes as outputs, illustrated in a study of Swedish pharmaceutical dispensing. Their method separates performance indices into quality, technical and efficiency change over time using a Malmquist productivity index, under assumptions of constant returns to scale. In principle, this method could also be adapted to consider negative quality attributes, as inputs into technology.

In Färe, Grosskopf and Roos (2001) the method of Färe, Grosskopf and Roos (1995) was extended to allow for consumer satisfaction for non-traded quality aspects of care, using the utility indirect input distance function of Shephard (1974), which

---

admission across hospitals is unchanged, unless quality of care is significantly increased.

models technology with a utility target. In this later study, in justifying use of consumer values in efficiency analysis, Färe, Grosskopf and Roos (2001) state:

> "We would argue that it is useful to consider consumer satisfaction in conjunction with the resource decisions of firms – especially in cases like the public sector where price signals from the consumer are unavailable or distorted." Färe, Grosskopf and Roos (2001:216)

If panel data, across hospitals, over time of admissions, physical inputs, and disutility events were available, then quality, technical and efficiency change could be estimated in hospitals, specifying disutility events as inputs using the general method of Färe, Grosskopf and Roos (1995). Where there are multiple quality aspects (types of disutility events), and relative preferences between non-traded quality inputs are available, they could be allowed for following the method of Färe, Grosskopf and Roos (2001). However, it should be noted that these methods do not allow quality change to be integrated into a combined measure with technical change and efficiency change. Technical change, efficiency change and quality change are calculated separately in each method, under the assumption that production is separable.

In considering performance, over time, policy makers may wish to consider technical change, efficiency and quality indices together, rather than partially under the assumption of separability. In hospitals, integrating quality and activity in determining performance, can be considered particularly important, given the derived nature of demand for hospital admissions. Cross sectionally, relative efficiency in production of admissions is not necessary for net benefit maximisation without conditioning on quality of care, as argued in chapter 2, and illustrated under the correspondence theorem in chapter 5. Similarly, over time, change in performance would ideally integrate changes in quality and resources per admission under a trade-off between the cost and value of quality of care.

If inputs, disutility events and their relative prices were available over time, then, under the correspondence theorem, a net benefit index of performance measurement could be constructed in minimising the sum of weighted inputs, including disutility events per admission (valued at WTP to avoid them). Change in performance, between any time periods, could be estimated using any appropriate index method that

allowed for inputs and disutility events as inputs (with associated relative factor prices and willingness to pay) and admissions as output. For example performance measured between periods *i* and *j* could be represented by a Fischer index of net benefit change over time as in equation 7.1

$$I_{i,j} = \sqrt{\left(\frac{y_j}{X_j \cdot W_i + DU_j \cdot K_i} \Big/ \frac{y_i}{X_i \cdot W_i + DU_i \cdot K_i}\right) \times \left(\frac{y_j}{X_j \cdot W_j + DU_j \cdot K_j} \Big/ \frac{y_i}{X_i \cdot W_j + DU_i \cdot K_j}\right)} \qquad (7.1)$$

Where:

$I_{i,j}$ represents a Fischer index of change in net benefit performance between periods *i* and *j*;

X_i represents a vector of inputs $(x_1,..,x_m)$ in period *i* and W_i a vector of their factor prices in period *i*;

DU_i represents a vector of disutility events[9] $(du_1,..,du_p)$ in period *i* and K_i the vector of decisions makers value for avoiding these events in period *i* and;

y_i represents admissions in period *i*.

7.4.3 Capitation payments, equity objectives and budget control

In many countries and regions, an overarching capitation funding system to regions, at a tier above distribution of provider payments, is current practice in use of case-mix payment systems (Rice & Smith, 2001). These capitation payment pools, based on expected (risk adjusted) resource needs of populations in regions, both reflect equity considerations in resource allocation and provide budgetary control.

In considering equity as an objective, capitation based funding to geographic areas attempts to provide equal opportunity of access to resources for services for equal need (risk adjusted) across regions. However, as capitation systems over-arch a DRG case-mix payment system to hospitals, equal access is to services in which providers have economic incentives, as agents, for quality of care minimising cost per

[9] If disutility events represent continuous variables such as limitation or disutility then disutility events correspond to the rate multiplied by number of admissions.

admission. In practice, this leads to quality of care that is variable across providers, depending on local ability to oppose economic incentives, for minimising cost per admission, inherent in case-mix funding.

An equity objective of equal access to equal quality of care for equal need, is likely to be better served by a system that provides appropriate incentives for net benefit maximising rather than cost minimising quality of care. Economic incentives consistent with evidence-based medicine, in health technology assessment and other regulatory mechanisms and objectives, provide a better chance of converging at a standard (as well as more appropriate) quality of care. Economic incentives for net benefit maximisation support, rather than (under case-mix funding) directly oppose, evidence-based quality of care in practice.

The hospital funding mechanism, illustrated in chapter 6, allows such a system, providing economic incentives for net benefit maximisation, where correspondence conditions are met. Under the proposed funding mechanism, a capitation model, based on relative need by region, can, as with case-mix funding, overarch payments to hospitals within regions in allowing for equity objectives and providing budgetary constraints. Adjustment of payments across hospital providers within a region for their relative quality of care (represented by disutility event rates), can occur within current case-mix payment per admission by DRG, as identified in equation 6.5 in section 6.3.

Where incentives for net benefit maximisation, rather than cost minimisation, are provided, conditions for equity in access to appropriate quality of care (and hence expected health outcomes), for equal need are created. It should be stressed that this argument for improved equity in the proposed funding mechanism is dependent on correspondence assumption being satisfied. In particular adjusting for differences in risk of disutility events and costs at a DRG level across hospitals given predictive factors across patient populations is required to prevent incentives for cream-skimming.

Adjustment of payments for risk factors within-DRG should already be a policy focus with case-mix funding, in overcoming incentives for cream-skimming based on

differences in expected costs of patients. Payments conditioning on disutility event rates for any given DRG, do, however, require additional adjustment for expected disutility event rates based on identifiable risk factors within-DRG, as described in section 6.8. In the absence of appropriate adjustment for identifiable risk factors, economic incentives can be created to cream-skim, denying access to those in greatest need. The challenge to policy makers, as described in section 7.2, is to model disutility event rates with patient level data at a DRG level to allow robust and appropriate adjustment for patient risk factors.

7.5 Supporting appropriate referral practices

Coverage of effects by disutility events requires either data linkage to effects beyond-separation, or measurement of health state at separation. At a DRG level, decision analytic methods can be used to identify costs, and effects, attributable to quality of care. Linked data for disutility event rates in hospital, and other health care settings (e.g. death registry, readmission) can be simply added as additional terms, given the linear nature of the correspondence theorem. In determining attribution of disutility events, disutility event rates can be adjusted for patient covariates at point of admission and, timing of events and environment post-separation, as described in section 5.7.2.

Data linkage using decision analytic methods can, therefore, be employed to include attributable post-care cost and events adjusted for risk factors and timing, to create economic incentives for appropriate referral practices. Under these conditions, it is only if hospitals expect a higher net benefit of treating in an alternative setting that they have an economic incentive to refer. This compares with economic incentives within case-mix funding, to always refer on to further care, abrogating responsibility of care to engage in cost-shifting.

7.6 Appropriate incentives from case-mix funding: a case of the Emperor's clothes?

Proponents of case-mix payment have made claims of case-mix funding such as:

> "It rewarded efficiency and growth in services while at the same time guarding quality" (Brook, 2002).

However, this argument of case-mix funding 'guarding', or providing appropriate incentives, for quality of care, is based on a fallacy of composition. For cost minimisation to be consistent with net benefit maximisation requires counterfactual assumptions that:

1. the infra-marginal cost saved from avoiding treatment of negative effects with high quality care always outweigh the intra-marginal cost of higher quality care across patients for all quality of care and;
2. negative effects of care are treated within the same hospitalisation.

As was illustrated in figures 5.1 and 6.1, the first assumption is violated theoretically, and in practice, with diminishing marginal returns increasing resources used to improve health outcomes, as the disutility event rates approaches 0 (quality of care increases). The second assumption is violated where providers have discretion with respect to point of separation in patient care, and effects of care take time to evolve. Provider discretion with respect to referral, or point of separation, provides scope to act on economic incentives to shift costs and effects of care beyond point of separation. To overcome the fallacy of the second assumption requires accountability for either health status at point of separation and costs beyond-separation, or effects of care (both health and cost effects) beyond-separation.

In including the costs, but not the value, of quality of care, economic incentives are created by case-mix funding to cost shift beyond-separation and lower quality of care below that of net benefit maximisation. This is particularly the case in case-mix funding where, as in the Australian State of Victoria:

> "The price paid for inpatient episodes is determined solely by the relative weight of the DRG." (Brook, 2002)

As quality of care has an effect on both cost (resource use) and value (effects) of care, cost minimisation per admission is not a sufficient, or necessary, condition for net benefit maximisation. The deliberate 'clinical neutrality' (Brook, 2002) of case-mix funding does not provide neutral economic incentives for quality of care. Rather, ignoring effects of care provides economic incentives for cost minimising quality of care and allows technical inefficiency to be hidden behind lower quality of care.

As outlined in chapter one, further claims made in support of such case-mix funding is that it:

> "… has enabled hospitals to make more informed decisions on best and most appropriate use of their resources. Case-mix funding encourages more efficient patient treatment and recognises the costs associated with different procedures." (Brook, 2002).

To have substance, claims of more efficient patient treatment in provision of health services, and guarding quality of hospital service provision, require inclusion and accountability for the value of effects and not just implicitly the cost of quality of care. The minimisation of costs (or resource use) conditional on quality of care, is a necessary, while not sufficient, condition for net benefit maximisation, as was identified in section 5.5.4 and illustrated in figure 5.1.

Payments based on expected costs conditional on quality of care do not, however, provide appropriate economic incentives for quality of care. Payments reflecting expected costs conditional on quality of care, while addressing technical inefficiency conditional on quality of care, do not provide incentives for net benefit maximisation. Rather, they allow bsses in net benefit attributable to allocative efficiency result, where quality is too high or low. Payments based on expected cost, given quality of care, are problematic, both because expected costs can increase with high rates of disutility events (lower quality of care), but also due to the lack of economic constraints this implies for improving quality of care as disutility event rate approaches 0.

Payments proposed under the correspondence theorem are conditional on effects of care relative to best practice. These payments are not adjusted for expected costs, conditional on effects of care, but rather the decision makers' value of quality of care relative to best practice. Hospitals, therefore, face a constraint requiring them to trade-off the cost and value of quality of care. Where correspondence conditions are satisfied, this trade-off is consistent with net benefit maximising quality of care. Under this mechanism, there are active economic disincentives for quality of care above, or below, a net benefit maximising level.

For protagonists of case-mix funding such as Brook (2002) who, despite the lack of a theoretical basis, maintain that case-mix payment have produced incentives for quicker, higher quality treatment and healthier patients, applying the performance and funding framework proposed under the correspondence theorem will 'reinforce' these incentives. Measurement of performance and funding will finally recognise, and reward, the value of higher quality of care.

However, in general, it should be clear to economists, that economic incentive to provide appropriate quality of care can only be created where hospitals are accountable for cost and effects of quality of care. Application of the correspondence theorem provides such accountability, with a trade-off consistent with evidence-based medicine in maximising net benefit, where correspondence conditions are satisfied. Explicit and more appropriate incentives for quality of care are created, corresponding with health care objectives by activity. The reliance of case-mix funding on partial performance measures, localised incentives from clinical standards or regulation, and local negotiation conditions for cost minimisation versus health maximisation within-hospital, to oppose economic incentives for quality of care minimising cost per admission, is removed.

7.7 Policy implications for health technology assessment (HTA)

7.7.1 Policy implications for translating HTA evidence into practice

A policy focus across health systems has been on translating evidence-based medicine from health technology assessment into practice in response in part to evidence of practice variation in the UK (NHS, 2000) the US (Marshall, Shekelle, Leatherman and Brook, 2000) and Australia (ACHCS, 2001; Fahey and Gibberd, 1995). However, case-mix funding and performance measures, in failing to provide an economic value for health effects of care, act as a barrier to translating evidence of net benefit maximising practices from health technology assessment into practice. Funding and economic performance measures that ignore effects of care provide economic incentives for cost minimisation, rather than net benefit maximisation (implicit in HTA decision making).

In attempting to create evidence-based medicine in hospital activities, it is important to note that health technology assessment with randomised control trial evidence has, in the main, been restricted to decisions with respect to new technologies. Evidence of net benefit maximisation is further restricted to trials where costs, as well as effects, of care have been collected or modelled. Adoption of evidence-based medicine in practice, therefore, clearly has gaps in providing appropriate incentives to choose net benefit maximising technologies across all hospital activities (DRGs).

Even where evidence allowing assessment of net benefit maximisation in trial settings is available, differences often exist between the setting of a trial protocol (other countries) and that of practice. These differences may typically include those related to characteristics of patient populations (particularly severity), difference in practice (strategy employed), the availability of complementary therapies, or relative prices of resources. The expected incremental effects, and resource use (costs), of care, in practice, may, therefore, be very different to evidence from trials.

Choice of evidence-based technologies, or strategies of care, also does not consider variation across providers, in the method or the proficiency with which, available technologies, or strategies of care, are adopted. This is particularly important to consider where there are economic incentives for provision of cost minimising quality of care in practice.

In comparison, where funding mechanisms and performance measures are consistent with maximising net benefit, natural incentives and feedback loops are created for providers to maximise net benefit in their choice and use of technologies in treating patient populations. Where correspondence assumptions are satisfied, application of the correspondence theorem allows a natural mechanism creating incentives from performance measurement and funding that reinforce net benefit maximising evidence-based medicine in practice. Net benefit maximising behaviour is, therefore, reinforced across a continuum from health technology assessment through to performance measurement and funding in practice.

7.7.2 Policy implications for health technology assessment decisions

Where health technology assessment indicates uncertainty in assessment of effects of care, high costs of reversing bad decisions act as a barrier to allowing providers freedom to adopt potentially valuable technologies. Technologies may not be as effective in practice, as evidence in (randomized control) trial settings suggests. Currently there is no systematic mechanism to monitor, identify or reverse such technologies which do not perform in practice. Hence, there are high costs of reversing decisions to support technologies and decision-makers need to be cautious where there is uncertainty, to avoid these high costs of reversal.

Identification, monitoring and verification of effects, under a performance measurement and funding framework based on the correspondence theorem, allows incentives for net benefit maximisation to be created, where correspondence conditions are satisfied. Under correspondence conditions, with performance and funding mechanisms valuing disutility events (quality of care) at the decision maker's value of WTP, each hospital effectively becomes an agent for evidence-based medicine, with economic incentives to maximise net benefit in choice, and use, of technologies. Consequently, if effects in practice are not net benefit maximising, technologies can be naturally reversed by incentives to adopt more cost-effective technologies in performance measurement and funding.

Under correspondence conditions, with funding and performance measurement conditional on effects of care, a continual feedback loop from relative net benefit in the adoption of a technology, in practice, are created. Control of the diffusion of uncertain technologies, therefore, becomes feasible in practice. In option value terms (Weisbrod, 1964; Arrow & Fischer, 1974; Henry, 1974; Pindyck, 1988; Pindyck, 1991), the expected cost of reversing decisions to adopt technologies which turn out to be ill-advised are reduced, if provider behaviour can be naturally reversed. Currently, in the absence of a natural reversal mechanism in funding, decisions with respect to new technology made under uncertainty are often economically irreversible, in the sense proposed by Bernanke (1983:86, bracketed text added) that: "once constructed they (decisions) cannot be undone or made into a radically different project without high costs", and clarified further by Tirole (1988:308) as: "the cost of being freed from the commitment … is sufficiently high that it does not pay to be

freed". To the extent that a natural monitoring and reversing mechanism reduces the cost of reversing decisions, the option value of delaying a decision with expected net benefit, but uncertainty, in order to wait for more evidence (and avoid potential costs of a bad decision), is reduced.

Therefore, where expected net benefit of a new technology is positive, but uncertain, the decision to adopt technology could be made sooner. Expected net benefit can be more readily exploited, where a natural monitoring and reversal mechanism is present, to reduce costs associated with reversing use of technology which, in practice, does not maximise net benefit.

7.8 Summary

The proposed performance method and funding mechanism, outlined and illustrated in chapters 5 and 6 respectively, provide policy makers with a systematic framework for including quality of care with an appropriate trade-off between value and cost of quality. The framework is not a panacea in itself. In ensuring correspondence with maximising net benefit, existing agendas for policy makers and researchers under case-mix funding remain in:

1. using patient level data to model risk factors for disutility events, and their costs, to enable adjustment in preventing incentives for cream-skimming (a challenge shared by capitation systems, insurers and case-mix payment) and;
2. enabling data linkage to prevent incentives for cost and event shifting.

These agendas can be systematically addressed at a dinical activity level using decision analytic methods to identify effects attributable to care. The linear framework, provided by the correspondence theorem, allows multiple disutility events, including linkage to events beyond-separation, and utility bearing aspects of care reframed as disutility events. To monitor that reporting of events are not being finessed, clinical audit or peer review processes can be employed in random samples of hospital patients.

The benefits of creating incentives for maximising net benefit, rather than cost minimisation, and avoiding cost-shifting and cream-skimming, in the measurement of

performance and funding of hospitals, are considerable. Under correspondence conditions, hospitals with these economic incentives, become agents of evidence-based medicine, maximising net benefit in:

1. choice and use of technologies;
2. systemic practices and;
3. referral practices.

Current economic incentives for reduced quality of care, hiding of inefficiency behind low quality of care and cost-shifting, can be systematically replaced by accountability for quality, and cost, of care in an evidence-based framework.

In considering the internal hospital organisation, negotiation between the two firms of medical service providers (clinicians) and administrators that Harris (1977) identified within hospitals, has been suggested to be more appropriate with an underlying framework of net benefit maximisation than cost minimisation. With payments conditional on disutility event rates representing effects of care, administrators can no longer act as accountants in attempting to minimise costs per admission under case-mix funding, but are rather required to trade-off the costs and value of quality of care in minimising cost plus value of disutility events per admission. Clinicians, while having their quality of care valued, become accountable for this quality of care. The primary source of hospital inefficiencies suggested by Harris (1977) in separation of clinicians with an objective of health maximisation and administrators with an objective of cost minimisation, characteristic under case-mix funding, is directly addressed.

Excess capacity reduced by administrators intent on cost minimisation under case-mix funding is valued where it effects health related quality of care, overcoming what Harris (1977) characterises as breakdown in negotiation and consequent hoarding of resources by clinicians defending excess capacity (quality of care) under uncertainty. The lack of information flows between clinicians and administrators and synthesis of information, in relation to relationships between cost and effects of care, are replaced by the need to address a trade-off between cost and value of quality of care. In resource allocation at a clinical activity level the proposed funding mechanism encourages a process akin to program budgeting and marginal analysis in trading-off

the value and cost at the margin relative to objectives and values of effects of care inherent in funding schedules.

Performance measurement, and funding, can thus reflect decision makers' objectives, and in conforming to objectives in health technology assessment, provide incentives consistent with evidence-based medicine in practice. Creating these incentives has also been shown to provide a natural mechanism for the identification, and reversal, of technologies whose expected effects from evidence in control trial settings do not eventuate in practice. Current high costs in identifying, and reversing, the use of non-performing technologies would be reduced, allowing decision-makers to adopt potentially valuable technologies, with positive net benefit, earlier.

The application of the correspondence theorem to performance measurement and funding, allows policy makers to create incentives consistent with net benefit maximisation under evidence-based medicine in health technology assessment. The strength of this consistency is further reinforced in chapter 8, where application of the correspondence method directly to health technology assessment is shown to have distinct advantages over current consideration of frontiers in the incremental cost effectiveness plane.

Chapter 8: The correspondence theorem applied in health technology assessment

8.1 Overview

The correspondence theorem, in providing an underlying objective function of maximising net benefit, allows a continuum between health technology assessment, performance measurement and funding, as described in section 7.6. The correspondence theorem can, however, also be directly applied to relative performance measurement in health technology assessment (HTA), where there are multiple technologies or strategies compared. In health technology assessment comparing multiple strategies or technologies, incremental consideration of performance with frontiers in the cost disutility plane is illustrated, with an example comparing colorectal cancer screening strategies, to provide advantages over current use of frontiers in the incremental cost effectiveness plane.

The correspondence condition of a common comparator is shown to be naturally met by randomised control trial evidence, while the coverage of effect by disutility event rate can be satisfied by reframing utility bearing effects. Rates of either survival, or reduction in morbidity, translate directly to disutility event rates, of mortality, and morbidity, respectively. Incremental life years or quality adjusted life years (QALYs) can be reframed relative to the most effective observed strategy as life years, or QALYs lost, respectively.

Distinct advantages, of incremental analysis relative to frontiers in the cost disutility plane, are shown to include the ability to:

1. measure extent of dominance with radial contraction;
2. provide a closed form with a statistic that does not rely on identification of a comparator;
3. illustrate differences in net benefit directly as levels of isocost curves.

A method for improving precision in bootstrapping of incremental cost effectiveness ratios, and bivariate distributions of costs and effects more generally, is also identified, and illustrated. Stratifying on prognostic factors, and undertaking ordering by prognostic risk in matching treatment and control re-sampled populations to form

replicates of the incremental cost effectiveness ratio or frontier of best practice, is illustrated to minimise structural uncertainty (inadvertently imposed with random matching of replicates).

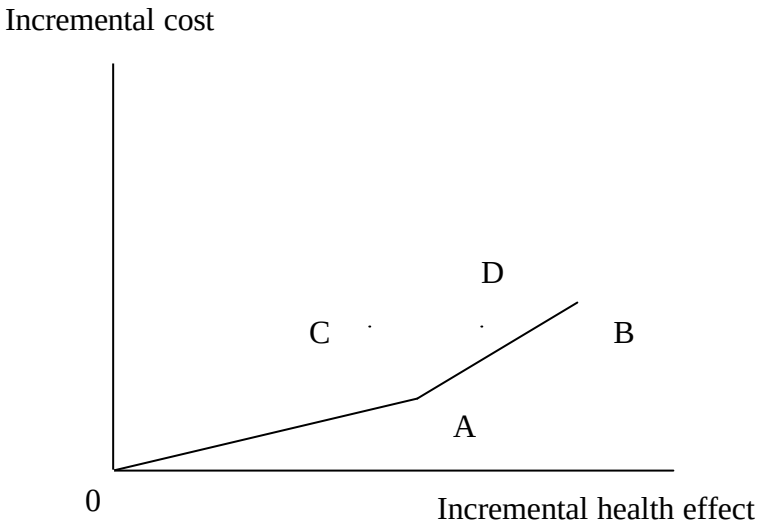
8.2 Comparing frontiers in the cost-disutility plane and the incremental cost-effectiveness plane

Currently, in identifying the frontier of potentially preferred technologies, or treatment strategies, for a given patient population, health technology assessment considers strategies in the incremental cost effectiveness plane. A comparator (usually representing current practice) is chosen as a reference point for axes in defining the incremental cost effectiveness plane. Potential best practice technologies, at given thresholds of willingness to pay, are defined by a frontier with south east movement in incremental cost effectiveness space.

Where another technology, or any linear combinations of alternative technologies has lower cost, and equivalent or higher effectiveness, this is reflected respectively in the terms 'dominance' and 'extended dominance' of a technology (Weinstein, 1990).[1] For example, in figure 8.1, in comparing strategies A,B,C and D, a frontier is formed by 0AB, with strategy C dominated by A, while a linear combination of strategies A and B has extended dominance over strategy D.

[1] The frontier can pass through the origin if the comparator strategy (usually defined as current practice) is not dominated. Otherwise the frontier passes below the origin, where a strategy or linear combination of strategies have lower expected cost at the same expected effect as the comparator.

Figure 8.1: Dominance and extended dominance in the incremental cost effectiveness plane



While dominance can be identified by these frontiers in the incremental cost effectiveness plane, they do not, however, have radial contraction properties. In comparison, if the correspondence theorem is applied to health technology assessment (HTA), then frontiers with radial contraction properties can be constructed in the cost disutility plane. In applying the correspondence theorem to comparison of relative technologies or strategies, the common comparator assumption translates to comparing technologies with patients facing the same baseline risk of costs and effects.

8.2.1 Satisfying the common comparator assumption in HTA

Where health technology assessment is based on randomised control trial evidence, the common comparator condition should be satisfied by randomisation (ideally stratified randomisation). If randomised populations are not 'balanced' across treatment arms, costs and effects need to be adjusted for prognostic factors in comparison. However, this requirement for adjustment would be the case whether in the incremental cost-effectiveness or cost-disutility plane.

Where health technology assessment is based on decision analytic modelling, the same baseline population risk of costs and effects of care should be used in applying relative treatment effects of strategies. The common comparator assumption should, therefore, be satisfied, by randomisation in the case of health technology assessment based on randomised trial evidence, and by the modelling of common baseline risk, to which estimated treatment effect by treatment strategy are applied, in the case of decision analytic modelling.

In general, if the common comparator assumption is satisfied in health technology assessment in the incremental cost effectiveness plane, it will also be satisfied in the cost disutility plane. The correspondence theorem can, therefore, be robustly applied if the second correspondence condition of disutility events covering effects of care is satisfied. That is, if effects in the incremental cost effectiveness plane can be mapped to rates of disutility events.

8.2.2 Satisfying coverage of health effects by disutility event rates in HTA

In relative performance assessment, if effects of care are measured for each strategy in the incremental cost effectiveness plane, they can be reframed as rates of disutility events. Where effects are measured as reduction in disutility event rates, such as mortality or morbidity, mapping of effect to disutility events is simple and direct. Incremental survival rate becomes mortality rate in the disutility plane, while reduction in the rate of morbidity translates to morbidity rate in the disutility plane. In the case of incremental effects, measured as life years or quality adjusted life years saved per person (patient), disutility event rates can be respectively constructed as life years or QALYs lost, relative to the most effective strategy or technology available.

Frontiers of best practice constructed in the cost-disutility plane, allow easy identification of dominance, and best practice technologies or strategies at given willingness to pay consistent with that in the incremental cost effectiveness plane. At any given willingness to pay, the optimum strategy, or convex combination of strategies, is where an isocost curve with slope of WTP is tangent to the unit isoquant frontier minimising cost and disutility events per admission. Dominated strategies in the cost-disutility plane are those with cost and disutility event rates both higher than

an alternative strategy, while extended dominance is where a strategy is dominated by a convex combination of other strategies.

In the cost-disutility plane, unlike the incremental cost effectiveness plane, dominated strategies targets and peer strategies can be explicitly identified with radial contraction to a frontier. In the incremental cost-effectiveness plane, the lack of a vertex prevents radial contraction. The direction in which dominated technologies should be compared in constructing the frontier can only be identified as somewhere south-east. Hence identification of the frontier cannot use methods with radial contraction such as DEA, rather requiring comparison of incremental cost effectiveness between linear combinations of strategies.

In comparison, in the cost-disutility plane, the ability to radially contract towards an origin using DEA allows a relative degree of dominance (technical inefficiency) to be estimated. Differences in net benefit are also easily identifiable, at any given threshold value for avoiding disutility events, as distances between isocost curves. These distances can be interpreted in either net monetary benefit terms, on the cost axis, or net health benefit or effectiveness terms, on the disutility axis.

8.3 Illustrating cost-disutility frontiers where effects are rates

To illustrate comparison between frontiers in the incremental cost effectiveness plane, and incremental comparison relative to frontiers in the cost-disutility plane, a simple hypothetical example is initially considered, with health effects measured as survival rates. Table 8.1 considers costs per patient and survival rates of 10 alternative strategies for a given patient population, where strategy 1 is assumed current practice.

Table 8.3.1: Average cost per patient and survival rate for ten hypothetical strategies

Strategy	1	2	3	4	5	6	7	8	9	10
Cost/patient	5000	5500	6500	7000	9000	10000	11000	12000	12000	12500
Survival rate	0.6	0.63	0.64	0.66	0.68	0.67	0.64	0.67	0.66	0.69
Mortality rate	0.4	0.37	0.36	0.34	0.32	0.33	0.36	0.33	0.34	0.31

For these 10 strategies and their costs and effects, figure 8.2 illustrates a frontier represented in the incremental cost effectiveness (survival) plane and figure 8.3 illustrates a frontier in the cost disutility (mortality) plane.

Figure 8.2: Frontier in the incremental cost-effectiveness (survival) plane

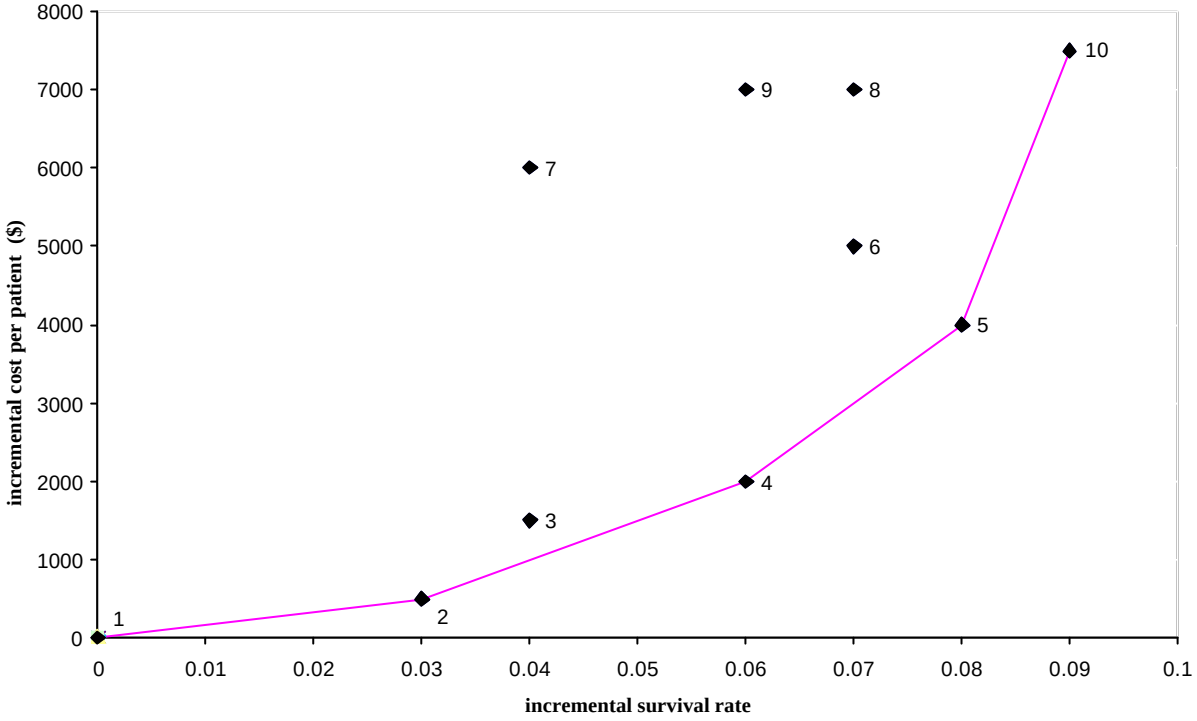
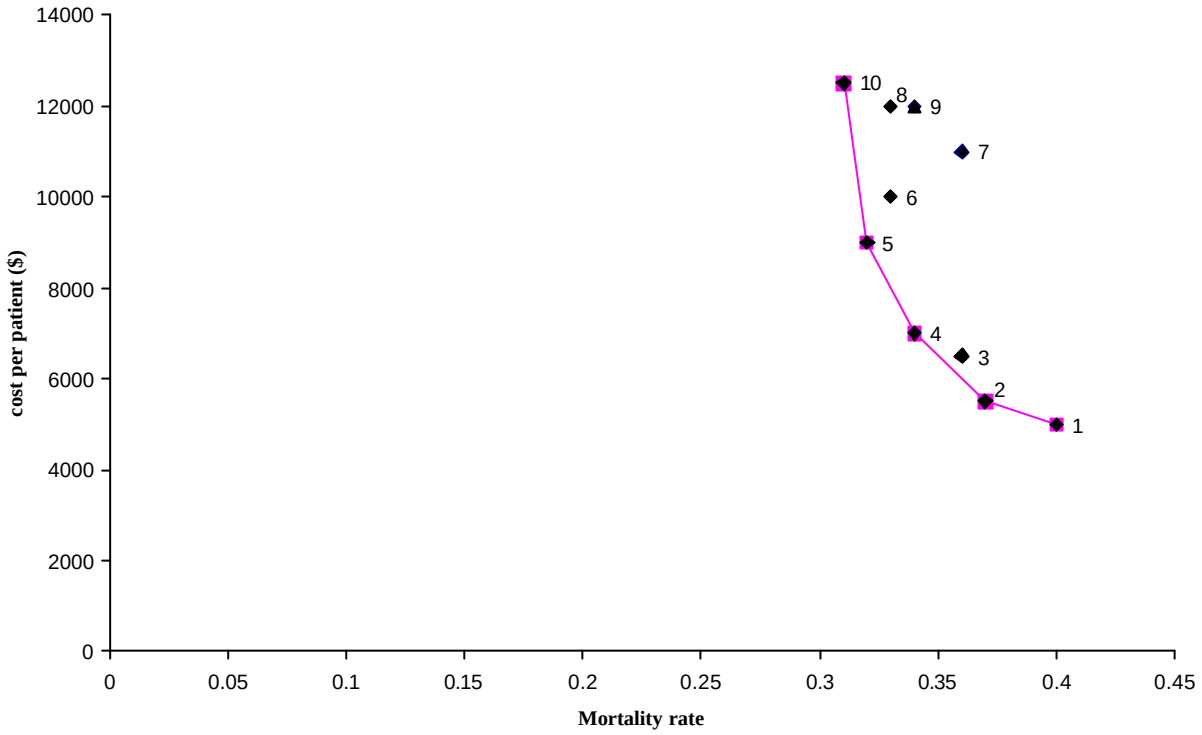


Figure 8.3: Frontier in the cost-disutility (mortality) plane



Both frontiers allow equivalent interpretation of dominance (strategies 6,7,8,9 are dominated) and extended dominance (strategy 3), although with radial contraction to an origin (reduction in cost and disutility event rates) this may be more intuitive in the cost disutility plane (figure 8.3). In the cost disutility plane, data envelopment analysis can be used to easily identify strategies on the frontier, with technically efficient strategies allowing no further radial contraction. Strategies on the frontier are those with technical efficiency, under constant returns to scale equal to one in table 8.3.2.

Table 8.3.2: Technical efficiency of strategies in the cost-disutility plane

Strategy	Technical efficiency under CRS	Cost per Patient ($)	Mortality rate
1	1.000	5 000	0.40
2	1.000	5 500	0.37
3	0.980	6 500	0.36
4	1.000	7 000	0.34
5	1.000	9 000	0.32
6	0.964	10 000	0.33
7	0.883	11 000	0.36
8	0.949	12 000	0.33
9	0.924	12 000	0.34
10	1.000	12 500	0.31

The use of DEA in the cost disutility plane also makes it clear that comparison of technologies, relative to the frontier, is under constant returns to scale. This assumption is also made, but not explicitly, in the incremental cost effectiveness plane. To identify regions of WTP threshold over which strategies on the frontier are best practice (maximise net benefit) in the incremental cost effectiveness plane, requires sequentially moving up the frontier, implicitly changing the axes for incremental cost effectiveness. In the cost-disutility plane, regions can be simply identified by back-solving for the threshold level of WTP between adjacent technically efficient strategies, ordered by cost or effect, using the same method described in section 5.6.5.

Using the cost per admission and mortality (disutility) rate of the technically efficient strategies, regions of threshold values over which strategies are preferred can be found

by back-solving on adjacent technically efficient strategies (ordered either by effect or cost) with equation (5.10):

$$C_i / y_i + (DU_i / y_i) \times k = C_j / y_j + (DU_j / y_j) \times k$$
$$\Leftrightarrow k = (C_j / y_j - C_i / y_i) / (DU_i / y_i - DU_j / y_j).$$

Table 8.3.2 identifies which strategies are potential best practice (technical efficiency =1). Regions of willingness to pay, over which these technically efficient strategies are preferred as best practice (economic efficiency is one), are shown in table 8.3.3, applying equation 5.10, with technically efficient strategies ordered by average cost per patient,.

Table 8.3.3 Willingness to pay regions of best practice for strategies ($ per life saved)

Strategy	Cost /patient ($)	Mortality rate	WTP region over which strategy preferred ($ per life saved)
1	5000	0.40	0 to $16,666
2	5500	0.37	$16,667 to $50,000
4	7000	0.34	$50,001 to $100,000
5	9000	0.32	$100,001 to $350,000
10	12500	0.31	$350001+

In the cost disutility plane, the question of preferred strategy in decision-making becomes a question of economic efficiency of alternative strategies at any given WTP, which can be indicated graphically by isocost curves tangent to the frontier. This is analogous to identification of best practice in figure 5.5, in section 5.6.1. For any given willingness to pay, distances between isocost curves values for alternative strategies represents differences in net monetary benefit per patient on the cost per patient axis, and differences in net effect benefit per patient on the disutility event rate axis. These differences are analogous to those in net benefit across providers illustrated in figure 6.5.

8.4 Cost-disutility frontiers where effects are life years or QALYs

Where effects of care are measured in life years or quality adjusted life years (QALYs), mapping from incremental effects to disutility space can be easily undertaken, by constructing disutility event rates relative to the most effective

strategy[2]. Years of life lost, or QALYs lost, can be measured relative to a standard of the most effective strategy.

This proposed method has similarities to the method employed in the Global Burden of Disease study (Murray and Lopez, 1996), in calculating life years and disability adjusted life years (DALYs) lost. However, Murray and Lopez (1996) used a threshold of highest attainable life expectancy (using life tables for Japan) for any given (age-sex) population. In comparing the effectiveness of strategies for a given patient population in health technology assessment, the most effective strategy observed can be used as a natural standard. Defining disutility relative to the highest observed effectiveness also has the advantage of preventing cost slacks in estimated technical efficiency scores, as the lowest disutility event rate on the frontier is then, by definition, 0.

8.4.1 Illustrating cost disutility frontiers where effects are life years

To illustrate use of frontiers in the cost disutility plane, where effects are measured by life years, a comparison of 22 strategies for colorectal cancer screening is undertaken, based on their expected costs and life years in 50 year old males at average risk of colorectal cancer. Table 8.3.4 summarises life expectancy and average cost per patient in $US (1998) modelled for each of 22 strategies for colorectal cancer screening reported by Frazier (2000).

[2] Mapping could also directly use disutility events if they were able to cover effects of care. In the case, valuing would entail separating populations by different combinations of disutility events and valuing relative to no disutility events, for each of these combinations, using decision analytic methods, as described in section 5.7.

Table 8.3.4: Expected costs and life expectancy of 22 screening strategies for colorectal cancer in 50 year old males at average risk

	Cost per patient ($US 1998)	Life expectancy (years)
No Screening	$1,052	17.3481
Sig1 @ 55y	$1,070	17.3632
Sig2 @ 55y	$1,095	17.3654
DCBE @ 55y	$1,200	17.3585
Sig1-10y	$1,218	17.3732
Sig2-10y	$1,288	17.3775
Col @ 55y	$1,312	17.3760
Sig1-5y	$1,438	17.3806
DCBE-10y	$1,514	17.3687
Sig2-5y	$1,536	17.3866
UFOBT	$1,584	17.3901
UFOBT + Sig1-10y	$1,804	17.4004
UFOBT + Sig2-10y	$1,810	17.4022
DCBE-5y	$1,872	17.3826
RFOBT	$1,986	17.3991
UFOBT+Sig1-5y	$2,023	17.4041
Col-10y	$2,028	17.3959
UFOBT+Sig2-5y	$2,034	17.4066
RFOBT+Sig1-10y	$2,214	17.4065
RFOBT+Sig2-10y	$2,226	17.4078
RFOBT+Sig1-5y	$2,428	17.4091
RFOBT+Sig2-5y	$2,448	17.4110

Sig1 = Sigmoidoscopy followed by colonoscopy in high risk polyps only
Sig2= Sigmoidoscopy followed by colonoscopy in any adenomatous polyp
DCBE= Double contrast barium enema
FOBT = Faecal occult blood test (UFOBT= unrehydrated, RFOBT=rehydrated)
Col = Colonoscopy
5y = every 5 years
10y = every 10 years
@ 55y = at 55 years of age

Figure 8.4 illustrates a frontier in the incremental cost effectiveness plane for these 22 strategies for colorectal cancer screening, based on the expected incremental costs and life years saved in table 8.3.4. Figure 8.5 illustrates a frontier based on the same data in the cost-disutility (life years lost relative to most effective strategy) plane.

Figure 8.4: Frontier of alternative screening strategies for colorectal cancer in the incremental cost effectiveness (life year) plane

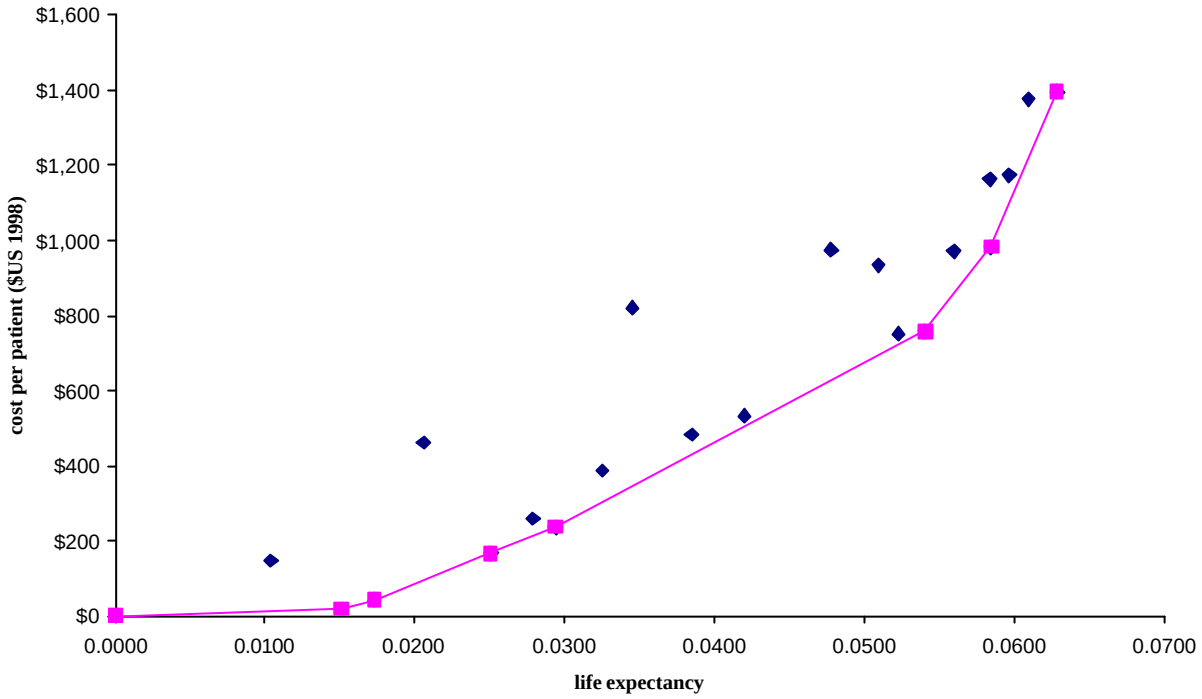
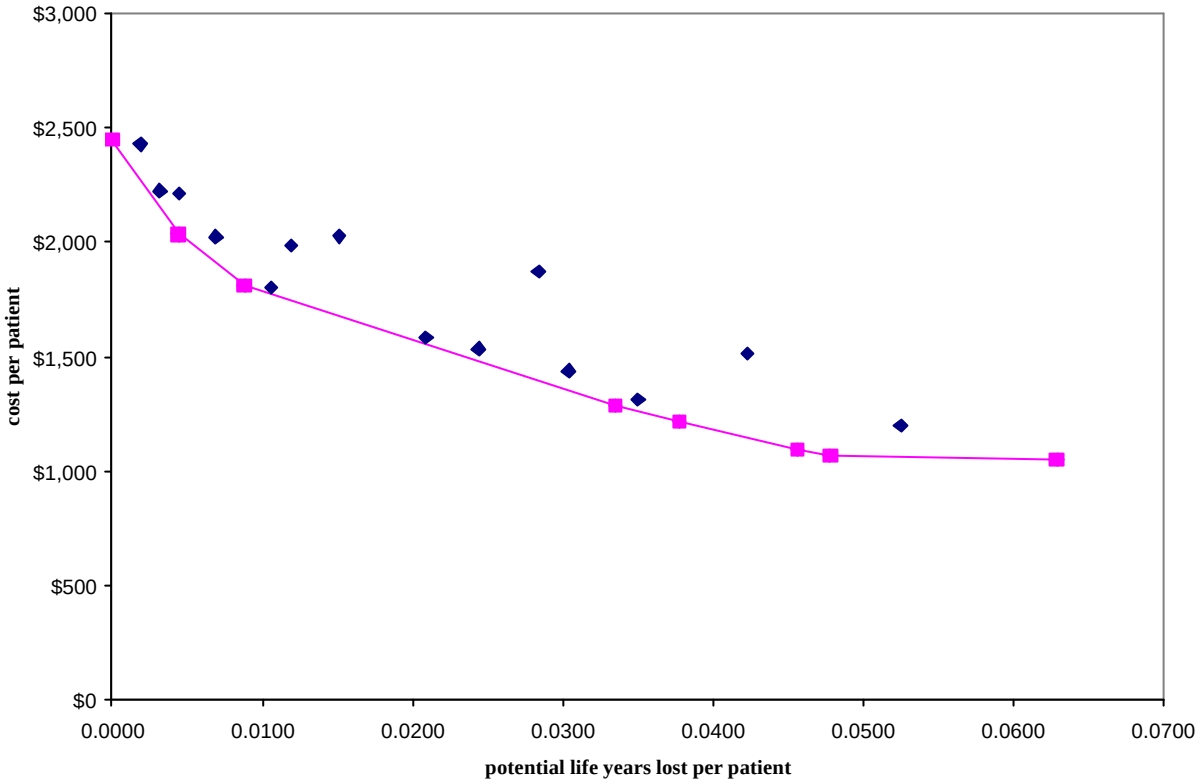


Figure 8.5: Frontier of alternative screening strategies for colorectal cancer in the cost disutility (life years lost) plane



Common regions of willingness to pay per life year saved over which strategies are preferred, are identified with relative performance measured in either the incremental cost-effectiveness plane or cost-disutility plane. In the case of the cost-disutility plane, DEA under constant returns to scale (CRS), with costs and disutility events as strongly disposable inputs, and admissions as a strongly disposable output, can be used to simply identify strategies on the frontier. Strategies on the frontier in the cost-disutility plane, are indicated by a technical efficiency of 1, or equivalently technical inefficiency of 0 in table 8.3.5

Table 8.3.5: Degree of dominance (technical inefficiency) of 22 screening strategies for colorectal cancer in 50 year old males at average risk

Strategy	*Technical efficiency*	*Degree dominated (technical inefficiency)*
No Screening	1.000	0
Sig1 @ 55 y	1.000	0
Sig2 @ 55 y	1.000	0
DCBE @ 55y	0.898	0.102
Sig1-10y	1.000	0
Sig2-10y	1.000	0
Col @ 55y	0.974	0.026
Sig1-5y	0.959	0.041
DCBE-10y	0.832	0.168
Sig2-5y	0.973	0.027
UFOBT	0.985	0.015
UFOBT + Sig1-10y	0.984	0.016
UFOBT + Sig2-10y	1.000	0
DCBE-5y	0.807	0.193
RFOBT	0.892	0.108
UFOBT+Sig1-5y	0.951	0.049
Col-10y	0.850	0.150
UFOBT+Sig2-5y	1.000	0
RFOBT+Sig1-10y	0.928	0.072
RFOBT+Sig2-10y	0.969	0.031
RFOBT+Sig1-5y	0.939	0.061
RFOBT+Sig2-5y	1.000	0

Sig1 = Sigmoidoscopy followed by colonoscopy in high risk polyps only
Sig2 = Sigmoidoscopy followed by colonoscopy in any adenomatous polyp
DCBE= Double contrast barium enema
FOBT = Faecal occult blood test (UFOBT=unrehydrated, RFOBT=rehydrated)
Col = Colonoscopy
5y = every 5 years
10y = every 10 years
@ 55y = at 55 years of age

While in the incremental cost-effectiveness plane the extent of dominance cannot be estimated, in the cost-disutility plane, dominated strategies technical (in)efficiency scores, allow estimation of the extent to which strategies are dominated (the last column in table 8.3.5).

Regions of willingness to pay for which strategies on the frontier (technical efficiency of 1 in table 8.3.5) are best practice (economically efficient) are simply calculated back-solving threshold values between technically efficient strategies ordered by cost

or effect. Regions over which screening strategies are preferred (economic as well as technical efficiency is one) are shown in table 8.3.6.

Table 8.3.6 Willingness to pay regions for preferred strategies ($ per life year saved)

Strategy	*Average cost per patient*	*Average life years lost relative to the most effective strategy*	*WTP region of preferred (NB maximising) strategy (1998 US$ per life year saved)*
No Screening	$1,052	0.0629	0 to $1,192
Sig1 @ 55 yrs	$1,070	0.0478	$1,193 to $11,363
Sig2 @ 55 yrs	$1,095	0.0456	$11,364 to $15,769
Sig1-10years	$1,218	0.0378	$15770 to $16,279
Sig2-10years	$1,288	0.0335	$16,280 to $21,336
UFOBT + Sig2-10y	$1,810	0.0088	$21,337 to $50,909
UFOBT+Sig2-5y	$2,034	0.0044	$50,910 to $94,090
RFOBT+Sig2-5y	$2,448	0	$94091 or more

At any given threshold value, differences in strategies net benefit per person can be represented as net monetary benefit or net effect benefit by the distances between isocost curves on the cost and disutility axes respectively. Differences in net monetary benefit and net life year benefit of strategies relative to best practice at $50,000 per life year saved are estimated in table 8.3.7.

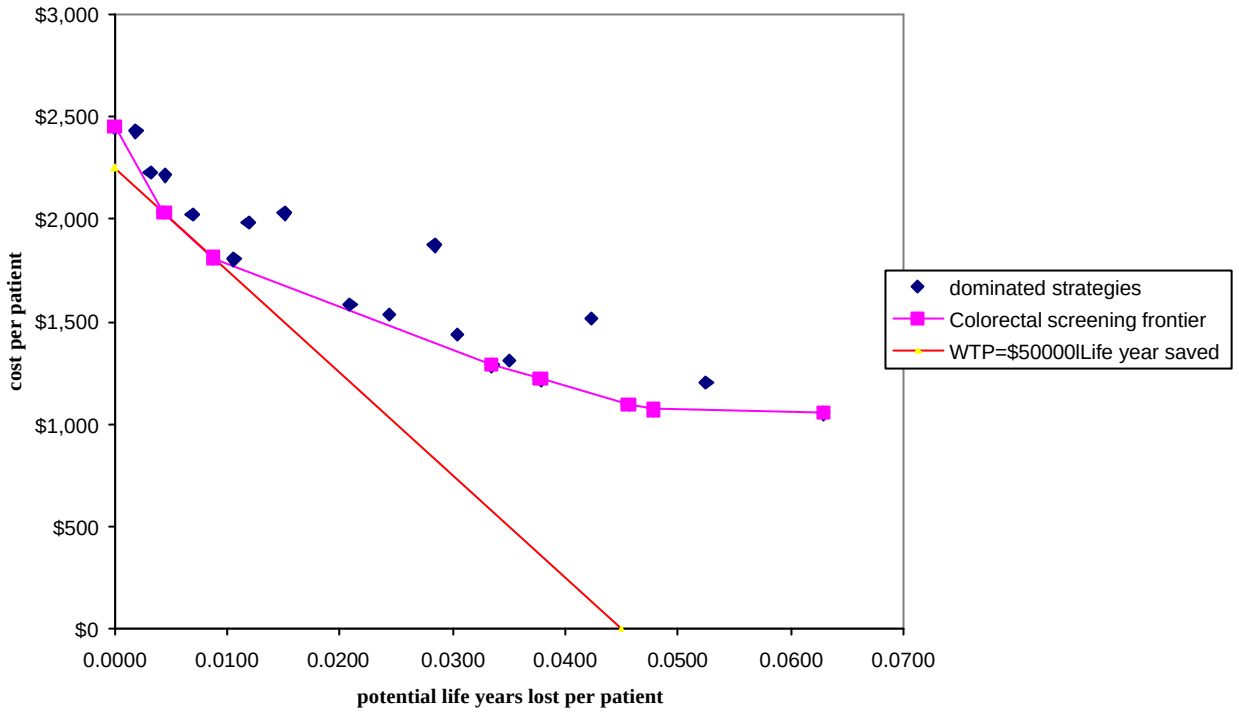
Table 8.3.7: Differences in net monetary and effectiveness benefit at $50,000 per life year saved of 22 screening strategies for colorectal cancer in 50 year old males at average risk

Strategy	*Loss in net monetary benefit per person (US$ 1998) at $50,000 per life year saved*	*Loss in net effect benefit (life years/1000) at $50,000 per life year saved*
No Screening	1,947	38.94
Sig1 @ 55 y	1,210	24.2
Sig2 @ 55 y	1,125	22.5
DCBE @ 55y	1,575	31.5
Sig1-10y	858	17.16
Sig2-10y	713	14.26
Col @ 55y	812	16.24
Sig1-5y	708	14.16
DCBE-10y	1,379	27.58
Sig2-5y	506	10.12
UFOBT	379	7.58
UFOBT+ Sig1-10y	84	1.68
UFOBT+ Sig2-10y	0	0
DCBE-5y	1,042	20.84
RFOBT	331	6.62
UFOBT+Sig1-5y	118	2.36
Col-10y	533	10.66
UFOBT+Sig2-5y	4	0.08
RFOBT+Sig1-10y	189	3.78
RFOBT+Sig2-10y	136	2.72
RFOBT+Sig1-5y	273	5.46
RFOBT+Sig2-5y	198	3.96

Sig1 = Sigmoidoscopy followed by colonoscopy in high risk polyps only
Sig2=Sigmoidoscopy followed by colonoscopy in any adenomatous polyp
DCBE= Double contrast barium enema
FOBT = Faecal occult blood test (UFOBT=unrehydrated, RFOBT=rehydrated)
Col = Colonoscopy
5y = every 5 years
10y = every 10 years
@ 55y = at 55 years of age

These estimates correspond to distances measured on the vertical and horizontal axes respectively between an isocost curve with a slope of $50,000 per life year saved, tangent to the frontier and isocost curves passing through each strategy. For example at $50,000 per QALY saved, Figure 8.6 shows the isocost line tangent to the frontier.

Figure 8.6: comparing net monetary benefit and net effectiveness benefit in the cost disutility (life years lost) plane



8.5 Advantages of HTA frontiers in the cost disutility plane

While dominance, and regions of willingness to pay, can be identified in either the incremental cost effectiveness plane, or cost disutility plane, comparing relative performance with frontiers constructed in the cost disutility also provides the ability to:

(1) estimate degree of dominance with radial contraction properties;

(2) intuitively represent decisions, including those with negative incremental costs and effect, relative to current practice with an isocost curve tangent to a frontier and;

(3) illustrate differences in net monetary benefit or net health benefit graphically and intuitively with levels of isocost curves.

8.5.1 Degree of dominance

In the cost disutility plane, technical inefficiency scores provide an estimate of the relative degree to which strategies are dominated. This is intuitive, given radial contraction unequivocally represents an improvement in both cost minimising and health maximising objectives, and compared technologies are as similar as possible in achieving these dual objectives. In the incremental cost effectiveness plane, it is unclear as to what direction dominance is focused. Measuring degree of dominance in the incremental cost-effectiveness plane, as a cost reduction or effect improvement, effectively ignores either costs or effects. Dominance can only be estimated in a common metric, as an absolute distance with a specified orientation in the incremental cost effectiveness plane, rather than as a ratio with radial contraction properties in the cost disutility plane.

8.5.2 Representing frontiers with negative effects relative to current practice

In relative comparison of technology, where strategies have negative effect relative to current practice (assumed to be the comparator), a threshold value of willingness to accept (WTA), rather than willingness to pay (WTP), for incremental health gain is appropriate. As described in section 5.7.3, loss aversion under prospect theory (Kahnemann and Tversky, 1979) suggests willingness to accept loss will be greater than willingness to pay for gain. This is supported empirically in health care, where evidence supports a 2-3 fold greater value for WTA health loss than WTP for

equivalent health gain (Willan O'Brien and Leyva, 2001). O'Brien, Gersten, Willan and Faulkner (2002) suggested this could be modelled in the incremental cost effectiveness plane with a kinked threshold at the current level of effectiveness (at the origin if this reflects current practice). Equivalently, in the cost-disutility plane this suggests a kinked isocost curve with slope WTP for effectiveness above (disutility rate below) that of current practice and WTA for effectiveness rates below (disutility rate above) that of current practice.

Figure 8.7 illustrates a hypothetical case for colorectal cancer screening where it is assumed that current practice is unrehydrated faecal occult blood test (UFOBT), willingness to pay for health gain is $50,000 per incremental life year saved and willingness to accept health loss is $100,000 per incremental life year lost.

Figure 8.7: kink in the threshold assuming UFOBT is current practice, with WTP=$50,000 per life year saved, WTA=$100,000 per life year lost

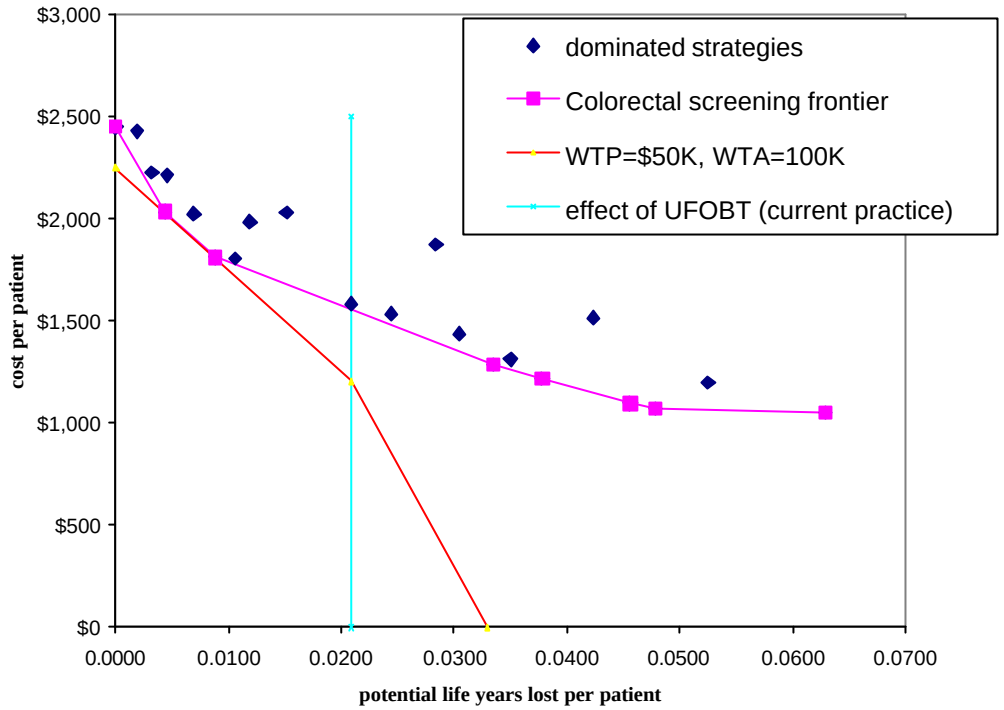


Figure 8.7 makes clear that a kink in the threshold, representing higher value of avoiding incremental loss (WTA) than gain (WTP), will, in general, make it less likely for strategies with reduction in effect (increased disutility event rates) to be preferred as best practice strategies. Convexity of the frontier, and concavity of the isocost curve representing decision making threshold values, also, however, clarifies that this kink cannot create problems of multiple solutions. The uniqueness of a preferred strategy or linear combination of strategies is not intuitively obvious in the incremental cost effectiveness plane.

The cost disutility plane allows the ability to characterise preferred strategies, given current practice, with an isocost curve kinked, at the level of effect of current practice, tangent to a frontier in the cost-disutility plane. This can be seen as an improvement on attempting to map preferred strategies in the incremental cost effectiveness plane, where incremental costs and effects may be negative relative to current practice. Where a strategy has negative incremental costs and effects, relative to current practice, it is unclear in the incremental cost effectiveness plane whether these strategies, or current practice, form part of the frontier. Currently, an approach to overcoming this problem is to redefine the origin as the strategy with lowest cost per patient, removing strategies that are dominated, and:

> "..comparing each option to the next more costly and more effective intervention." (Briggs, Goeree, Blackhouse and O'Brien, 2002:293).

This approach ensures, as with use of DEA in the cost-disutility plane, that the frontier remains convex, regardless of what is the usual current practice comparator. The decision making threshold can then, as in cost disutility space, characteristically have a slope of WTP above current level of effectiveness (below current rate disutility) and WTA below (above current rate of disutility). However, current practice will not necessarily form part of the frontier in the incremental cost effectiveness plane, or with the approach of Briggs, Goeree, Blackhouse and O'Brien (2002), be at the origin. Rather, current practice acts as a marker for the kink in the threshold, just as it did for a kink in the isocost curve in the cost disutility plane in figure 8.7.

8.5.3 Representing net benefit

When willingness to pay (WTP) to improve effects (reduce disutility rates) is known, differences in net benefit per admission, expressed as net monetary benefit per patient or net effectiveness benefit, can be represented by levels of isocost curves in the cost-disutility plane. These distances can be interpreted in monetary terms on the cost axis or effectiveness terms on the disutility event axis. Importantly, the isocost curves have a closed form of: cost per admission plus disutility events, valued at a constant willingness to pay equals a constant; which, when minimised, is equivalent to maximising net benefit, given correspondence theorem conditions are satisfied in HTA.

In the incremental cost effectiveness plane, distances between levels of lines with the same slope as the threshold value could also be graphically measured and characterised as distances in net benefit. However, levels of such curves are problematic, in not necessarily being able to be compared in the north east quadrant. Unlike isocost curves in the cost disutility plane, levels of threshold curves in the incremental cost effectiveness plane do not have a closed mathematical form for calculating differences in net benefit. They are not bound by axes, and their form of incremental effect, valued at WTP less incremental cost equal to a constant, requires comparison with an intermediate comparator.

8.6 Modelling uncertainty in the cost disutility plane

In undertaking health technology assessment, uncertainty of relative strategies' performance plane can be estimated, either with Monte-Carlo simulation, in the case of decision analytic models, or with bootstrapping, in the case of randomised control trials using patient data on cost and effects of treatments. In the cost-disutility plane, uncertainty of the degree, as well as the probability, of dominance can be estimated.

Appendix 8.1 identifies, and illustrates, a method to improve precision in bootstrapping the sample distribution of incremental cost effectiveness ratio. Stratified bootstrapping and ordering matching on prognostic risk factor score in re-sampled populations by treatment arm are shown to minimise structural uncertainty, inadvertently introduced by random matching of re-sampled treatment populations.

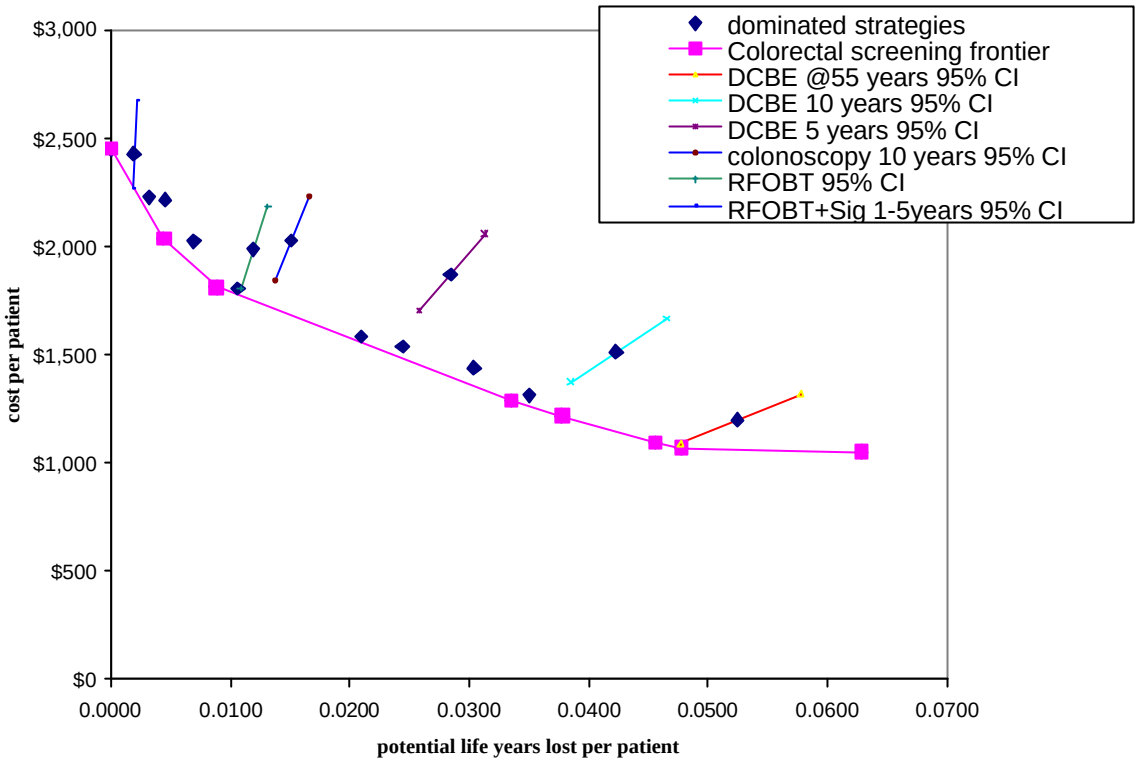
This method can be generalised to improve precision in bootstrapping of other statistics, based on bivariate distributions of cost and effects in comparing performance of technologies from randomised control trial patient data. In particular, the method can be used to improve precision in bootstrapping the uncertainty of dominance, where a strategy is compared with other strategies, to see whether it lies on the frontier or not.

8.6.1 Modelling probability of dominance in health technology assessment with Monte-Carlo simulation

In Briggs, Goeree, Blackhouse and O'Brien (2002), a Bayesian approach to estimating probability of dominance for any strategy, using Monte-Carlo simulation, was identified. Under this method, for each variable specified in a decision analytic model, random samples are taken from distribution of parameters for transition probabilities and outcomes (costs and effects) to reflect their 2nd-order uncertainty (Stinnett and Paltiel, 1997). A frontier is then constructed for each simulation to identify whether any strategy is dominated, or lies on the frontier. The proportion of simulated frontiers, in which the strategy is dominated, allows estimation of the probability of dominance for any strategy, given uncertainty of costs and effects.

In applying the correspondence method to health technology assessment in the cost-disutility plane, this Monte-Carlo simulation approach to estimating probability of dominance could be extended, to allow confidence intervals for degree of dominance to be estimated. Such confidence intervals could be graphically characterised using the property of radial contraction in the cost-disutility plane. Confidence intervals for degree of dominance (technical inefficiency) can be characteristically represented with radial contraction to, and away from, the frontier. Under uncertainty, a dominated strategies simulated (re-sampled) costs and effects can move relative to the frontier in any direction. However, radial contraction allows comparison with the closest comparators at baseline and is consistent with the frontier and its convexity properties. Figure 8.8 illustrates indicative radial ninety-five percent confidence intervals for probability of dominance, in comparing dominated strategies, relative to a frontier in the cost-disutility plane.

Figure 8.8: Indicative 95% radial confindece intervals for selected dominated strategies in the cost-disutility plane



8.6.2 Probability of dominance in health technology assessment estimated bootstrapping patient data

A bootstrapping approach can also be considered in estimating uncertainty of dominance, where patient level data on costs and effects from randomised trials, across strategies, is available. Following the bootstrapping approach described in Briggs, O'Brien and Blackhouse (2002), with patient level data, the sampled incremental cost-effectiveness ratio distribution can be estimated, re-sampling with replacement patients (to retain covariance structure between costs and effects for each patient) in each arm of the trial. This method then effectively randomly matches re-sampled treatment and control patients in forming incremental cost effectiveness ratio replicates. With patient level data on costs and effects for multiple strategies, each strategy's bivariate distribution in cost disutility space, but also the probability or degree of dominance, could similarly be bootstrapped, retaining covariance structure with random matching of replicates.

The precision of bootstrapping with random matching of re-sampled populations from treatment arms, implicit in the approach of Briggs, O'Brien and Blackhouse (2002), can, however, be improved upon. As described, and illustrated, in appendix 8.1, following Eckermann and Kirby (2003), structural uncertainty can be inadvertently introduced by random matching of re-sampled populations between treatment arms.

In modelling using Monte Carlo simulation, such structural uncertainty can, and should, be easily removed, using common baseline risk across strategies. Baseline risk of disutility events can have a distribution across replicates, but should be common for each strategy, compared within a replicate of the frontier.

Differences in baseline risk in randomly matched replicates, while not representing relative treatment effects, are predictive of incremental cost effectiveness ratios and degree of dominance. In bootstrapping, structural uncertainty can be minimised, using stratified re-sampling and ordering matching between re-sampled populations for treatments on prognostic risk of cost and effects of care, in forming replicates for comparison of strategies. This method of stratification, and ordered matching of replicates on baseline risk in bootstrapping, minimises structural uncertainty, without imposing any restrictions on analysis. This same general approach of stratification,

and ordered matching, could be used to minimise structural uncertainty in bootstrapping to estimate confidence intervals for degree, or uncertainty, of dominance.

Populations can be stratified on prognostic factors in re-sampling with replacement (bootstrapping) of their associated costs and effects for each strategy. These re-sampled populations for each strategy can, then, be ordered by prognostic risk, and matched on that order, in determining each replicate of the frontier. Using this method allows the probability of dominance from bootstrapping to be estimated, minimising structurally imposed uncertainty. This should be reflected in greater precision of confidence intervals for degree of dominance for any strategy, with the likelihood minimised of dominated strategies spuriously appearing on the frontier as an artefact of baseline risk, rather than treatment effect. In estimating uncertainty of dominance relative to a frontier, the importance of minimising such structural uncertainty increases, given comparison with all other strategies, rather than just the two strategies implicitly compared with incremental cost effectiveness ratios.

8.7 Summary

While the correspondence theorem can be applied to identify relative performance and funding consistent with net benefit maximisation in health technology assessment, it also can be applied in health technology assessment directly. Relative performance measurement with frontiers in cost-disutility space, using the correspondence theorem, has been demonstrated to have distinct advantages over frontiers in the cost-disutility plane. The common comparator assumption of the correspondence theorem is satisfied in health technology assessment with randomised trial evidence. Coverage of effects by disutility events is also naturally satisfied where effects are measured as rates (survival or morbidity), or in using life years or QALY lost relative to the most effective therapy.

Modelling frontiers for health technology assessment in the cost disutility plane allows:

1) Easy and intuitive identification of dominance of technologies or strategies where strategies are not on the frontier (technical efficiency less than1), and

estimation of the extent of dominance with the degree of radial contraction possible (technical inefficiency);

2) Regions of willingness to pay for adjacent technically efficient technologies, or strategies on the frontier, in back-solving for adjacent strategies on the frontier (technically efficient strategies ordered by cost);
3) Identification of appropriate comparators for dominated strategies in radially contracting cost and disutility event rates to best practice frontier. Currently, in the incremental cost effectiveness plane, it is unclear as to whether comparators for dominance are those south, east or southeast;
4) Relative differences in net benefit (Stinnett and Mullahy, 1998) to be identified at any given WTP from comparing isocost curves on which strategies lie (as in section 6.6), measurable as equivalent net monetary benefit on the vertical axis, and net health benefit on the horizontal axis and;
5) Uncertainty of dominance to be calculated, and indicatively represented with confidence intervals.

Uncertainty can be estimated either with Monte-Carlo simulation, where estimates are based on decision analytic models (Briggs, Goeree, Blackhouse and O'Brien, 2002), or bootstrapping, where patient level data has been collected in randomised control trials (Briggs, O'Brien and Blackhouse, 2002). Undertaking bootstrapping with stratification and matching ordering replicates on population's risk by treatment arm, improves precision of estimated confidence intervals for incremental cost effectiveness ratios, as illustrated in appendix 8.1. Where there are multiple strategies, this general method can also be applied to improve estimates of probability, and the precision of degree of dominance, in the cost-disutility plane.

Additional research in applying the correspondence theorem to performance measurement in other sectors, and methods for handling uncertainty in the absence of patient level data, are further explored in Chapter 9.

Chapter 9: Future research and extensions

9.1 Overview

Applying the correspondence theorem to performance measurement and funding in public hospitals has been illustrated to allow inclusion of quality of care at a clinical activity level, under a trade-off between the value and cost of quality. To allow trade-offs between value and cost of quality, consistent with net benefit maximisation, and to avoid incentives for cost-shifting and cream- skimming, requires satisfaction of correspondence conditions. To meet correspondence criteria at a clinical activity level in practice, a four-stage process is outlined:

1. identifying effects with decision analytic methods;
2. measuring these effects including data linkage (where required);
3. reframing them as disutility events, and;
4. adjusting for risk factors.

This four-stage process enables robust application of the correspondence theorem to performance measurement in practice, with a systematic evidence-based approach to satisfaction of correspondence conditions.

In applying the correspondence theorem to relative performance measurement, further research is suggested in allowing for the effects of sample variation. Parametric options suggested include:

1. regressing directly on cost plus disutility events conditional on WTP analogous to regression on the linear net benefit statistic outlined by Hoch, Briggs and Willan (2002) and;
2. stochastic frontier analysis in the cost disutility plane.

Hierarchical models with Bayesian methods of shrinkage estimation are also suggested as an avenue for future research. This follows application of such methods in allowing for sampling variation in comparing hospital quality of care indicators (Coory and Gibberd 1998; Simpson et al., 2003; Howley and Gibberd, 2003).

In allowing for uncertainty in hospital funding mechanisms, while aggregation across activities reduces the role of chance, risk sharing is suggested as an option particularly for smaller hospitals. In general, provided correspondence conditions are satisfied,

with differences in risk and effects beyond-care adjusted for, only hospitals with technical inefficiency hidden behind lower quality of care need fear the proposed funding mechanism.

Application of the correspondence theorem in performance measurement is suggested as appropriate in other service industries, where disutility-bearing event rates can be used to represent quality of care and maximising net benefit per service is an appropriate objective. In the public sector, natural applications are suggested in service industries such as corrective services, job placement, maintenance, quality assurance, safety information technology and security. As in hospitals, to avoid incentives for cream skimming and cost shifting in these settings requires the ability to adjust for risk factors prior to service and linkage to effects beyond point of service respectively. The linear framework provided by the correspondence, in combination with decision analytic methods, allows a framework for this to occur in a systematic evidence-based manner. The ability to reframe measures of utility, functional limitation or verifiable standards as disutility event rates, suggests the potential for wider application still.

9.2 Overcoming limitations in applying the correspondence theorem

Newhouse (1994), in critiquing the use of frontier methods at an aggregate level in estimating efficiency of hospitals, raised concerns of the ability of frontier approaches to allow for heterogenous outputs and include quality of care. Implicitly these concerns relate to the appropriateness of level of analysis and objective functions. By identifying an appropriate level of analysis and a frontier method consistent with net benefit maximisation, heterogeneity of outputs and health related quality of care have been addressed in this thesis. A DRG level of analysis allows problems of heterogeneity of output to be addressed, as well as avoiding the Fox (1999) aggregation paradox, as described in chapter 3. A clinical activity level also allows attributable quality indicators to be included in performance measurement.

In applying the correspondence theorem to allow an underlying objective consistent with net benefit maximisation, conditions of coverage of effects of care by disutility events and a common comparator are required to be satisfied. These conditions were

shown, in chapter 8, as able to be naturally satisfied in application to health technology assessment based on randomised control trial evidence. Robust application of the correspondence method to relative performance and funding of providers at a clinical activity level requires appropriate data measurement and adjustment to satisfy correspondence conditions.

9.2.1 Satisfying correspondence conditions in practice

In applying the correspondence theorem to performance measurement and funding, illustrations in chapters 5 and 6 were limited by the lack of linkage to effects beyond-separation and patient level data to allow adjustment for expected differences in cost and effects of care.

Therefore, while illustrations were instructive in identifying the ability of the correspondence theorem method to allow hospital performance measurement and funding consistent with net benefit maximisation, data linkage to effects beyond-care and patient level data are required for this to be robustly undertaken. The correspondence theorem is, therefore, not a panacea by itself. Existing challenges remain in linking to effects beyond-separation to overcome incentives for cost shifting, and risk adjustment to prevent incentives for cream-skimming, as outlined in chapter 7. As these challenges need to be faced in any performance measure, robustly allowing for quality of care, the explicit nature of these challenges, in satisfying correspondence conditions, is appropriate.

These challenges are focused by the systematic framework the correspondence theorem provides when applied with decision analytic methods at a clinical activity level. Meeting these challenges, under this framework, avoids cost shifting and cream skimming incentives, but unlike other frameworks, also allows incentives for net benefit maximising quality of care in practice.

The condition of coverage of effects of care can be satisfied using decision analytic methods at a clinical activity level in flexibly identifying effects of care. The linear nature of the correspondence theorem, and underlying net benefit framework, allows simple inclusion of multiple effects. These effects can include perceived utility

bearing aspects reframed as disutility-bearing event rates and effects beyond-separation with data linkage.

Adjustment for within-DRG differences, in expected risk and costs of patient populations, can allow satisfaction of the common comparator condition. Where the correspondence theorem is applied using data envelopment analysis (DEA), options in adjustment for differences include prior standardisation, inclusion of risk factors as non discretionary outputs and second stage regression. If expected disutility event rates (including that beyond-care with data linkage) are available from logistic regression, such as that of Simpson et al. (2003), their use as a non- discretionary output variable is suggested. This method restricts comparison to a linear combination of hospitals with equivalent, or greater, expected disutility event rates. The same approach could also be used to allow for differences in expected costs per patient due to patient risk factors. Including two non-discretionary variables of expected costs and disutility events does not face the restrictive comparison problems faced by Morey, Fine, Loree, Retlaff-Roberts and Tsubakitani (1992), where 12 such non-discretionary variables were employed.

Probabilistic data linkage to effects of care, such as that already undertaken in Western Australia (where hospitalisations can be linked to readmission, use of Medicare services, nursing home and mental health service as well as death and cancer registry), suggest risk adjustment and data linkage can be undertaken to allow satisfaction of correspondence theorem assumptions.

Using decision analytic methods at a clinical activity level, clinicians, epidemiologists and health economist can identify clinical and cost effects of care, as in health technology assessment (Weinstein and Fineberg, 1980; Drummond et al., 1987; Pettiti,1994; Drummond, 1997; Hunink et al, 2001), including those beyond-separation. Logistic models used at a clinical activity (DRG) level can adjust for differences in patient risks of clinical effects, and econometric modelling for differences in expected costs.

In allowing policy makers to robustly apply the correspondence theorem in practice, for each clinical activity (DRG) a four-stage process to satisfying correspondence conditions in performance measurement is suggested:

Stage 1 Using decision analytic methods to identify effects of (quality of) care, including beyond-separation effects (to prevent incentives for cost and effect shifting), to reflect objectives of care. Effects should include clinical effects such as morbidity, mortality, functional status at separation and resource effects such as readmission or use of other health care services.

Stage 2 Measure effects of care, including effects beyond-separation identified in stage 1, with data linkage where necessary. Linkage can be focussed to include variables such as re-admission by specific DRG, death registry data by specific cause, nursing home use by reason for admission, as well data for specified types of primary and secondary care services (Health Insurance Commission data in Australia).

Stage 3 Frame measured clinical effects of care as disutility event rates. Such disutility event rates can include mortality, morbidity, iatrogenic events, reduction in functional status measured on a cardinal scale, not meeting a standard of care, and disutility directly.

Stage 4 Adjust for differences in patient risk factors across hospitals to prevent incentives for cream skimming, and for other institutional, or post-separation environmental factors. A patient risk algorithm for each DRG is required to allow for differences in expected patient risks of disutility events and costs (already required in case-mix funding).

These stages allow an evidence-based approach to relative performance measurement in which, in valuing and creating accountability for quality of care under an appropriate trade-off with cost of care, policy makers address allocative and technical inefficiency.

9.2.2 Adjusting for environmental effects

Where external environmental constraints on hospitals exist, such as those related to size of operation (patient catchment area), access to technology, prices of factor inputs or differences in objectives, then they can also be allowed for in relative performance measurement.

Constraints of size of operation can be allowed for by using variable as well as returns to scale formulations for DEA, as described in 5.6.3. Fixed inputs (such as capital in the short run) can also be modelled as non-discretionary variables (Kopp, 1981; Banker and Morey 1986b), and can be separated from discretionary inputs which are radially contracted in input-orientated DEA models. Ordinal *a priori* differences in patient risk by hospital type can be allowed for, using the peer grouping method of Banker and Morey (1986a), as illustrated in section 5.6.6.

Where differences in access to technology across hospitals are determined externally, for example availability of capital equipment in principal referral but not other public hospitals, comparison can be undertaken within, and between, peer groups based on access to technology, as illustrated in section 5.7.1.2. While this method allows net benefit maximising best practice, as well as average performance of providers in using different technologies, differences in patient populations treated need to be considered.

9.2.3 Use of inputs rather than cost data

The use of cost, rather than factor, input data in the net benefit framework may be seen as a limitation for performance measurement, particularly over time, or where prices differ across providers due to exogenous factors. The extent of this limitation is, however, no greater than that of current measurement of performance with cost per case-mix adjusted separation. However, if factor input data were available, technical efficiency could ideally be calculated in radially contracting variable factor inputs and disutility events to a frontier of best practice, as section 5.5.4 described.

9.2.4 Allocation of joint costs

In measuring performance at a DRG level, a practical limitation is the allocation of joint costs down to a DRG level, within a consistent accounting framework. In

Australia, an accounting framework already exists, where since 1995 cost data at a DRG level across hospitals has been routinely collected by the National Hospital Cost Data Collection (Commonwealth Department of Health and Aged Care, 2000), in constructing Australian National DRG case-mix weights. This data was used in illustrating performance measurement and funding at a clinical level, in chapters 5 and 6, demonstrating the problems of incentives created by case-mix funding and performance measurement.

9.2.5 Theoretical limitations of a net benefit framework

Questions have been raised, in health technology assessment, relating to the assumption of a constant threshold value of willingness to pay for effects. Under correspondence conditions, this translates to a constant value of avoiding disutility events[1]. In section 5.7.3, the issue of higher value for willingness to accept, versus willingness to pay was considered, following prospect theory (Kahnemann and Tversky, 1979). This was suggested to further reinforce the need for inclusion of a value for quality, as a 0 value for quality provides incentives for quality of care which risk health losses relative to an endowment or entitlement (which may be considered best alternative care).

O'Brien, Gersten, Willan and Faulkner (2002) demonstrated that in the incremental cost effectiveness plane, where the origin represents current practice, a higher value of WTA, versus WTP, can be represented by a kink at the origin in modelling cost effectiveness acceptability threshold regions. However, they also allude to potential curvature in these regions, suggesting:

> "It might be questioned whether thresholds in either quadrant are straight lines or whether utility diminishes with greater quantity of program benefit." (O'Brien, Gersten, Willan and Faulkner, 2002:179)

In considering such curvature, Birch and Gafni (1992) suggest that the opportunity cost of additional resources needs to be accounted for in health technology assessment decisions. However, this is not allowed for with a constant shadow price, represented

[1] As section 5.7.2.6 described, decision analytic methods could be employed to identify the value of avoiding disutility events, with a common metric such as quality adjusted life year QALYs. This allows a robust method for valuation in maintaining a one-to-one correspondence with net benefit.

by a threshold value of willingness to pay. In health technology assessment, the shadow price for health gain is unlikely to be independent of the size of a health program considered, as the net benefit framework assumes. A constant WTP does not allow for increasing opportunity cost of additional resources in the case where technologies are more expensive and more effective. Birch and Gafni (2002), consequently argue that using a net benefit framework, with a constant threshold value of willingness to pay in health technology assessment decisions, may lead to unrestrained expenditure and, implicitly, short changing of the 'bang for the buck' that such approaches claim.

The argument of Birch and Gafni (1992 and 2002) for consideration of opportunity costs is clearly a valid criticism, of applying a constant threshold ratio, when considering decisions related to health technology assessment. However, while an argument for curvature of the acceptance region in health technology assessment, to account for opportunity cost, has some merit, this does not necessarily translate to considering relative technical efficiency in hospitals.

The abstraction of a common value for health gain, across hospitals, is more reasonable in considering relative efficiency of hospitals. There is no necessity for resources to increase from current case-mix funding levels, under the sequential two stage funding mechanism, described in chapter 6. In making hospitals accountable for effects of quality of care, the industry budget per admission for any given DRG across hospitals, currently allocated on the basis of average cost per admissions, is not proposed to change.

Rather than increasing resources, it is removing the ability to hide technical inefficiency behind low quality of care, with accountability for quality of care relative to best practice, driving health gain. The effects of improved quality on the health system over time, in response to accountability for quality, may even suggest potential for reduction in costs of treating healthier patient populations over time (with the potential exception of where quality reflects reduction in mortality rates).

9.3 Future research allowing for uncertainty

In allowing disutility event rates to be interpreted as quality of care indicators for hospitals, patient case-mix can be adjusted for, as described in section 5.7.1, clinical audit can potentially control for measurement bias in assessing disutility events and data linkage allow for effects beyond-separation. However, allowing for sampling variability is more problematic.

Sample variation, attributable to stochastic uncertainty in patient's costs and effects, can result in a more optimistic best practice frontier than underlying best practice, in estimating a best practice frontier with DEA. In the cost disutility plane, hospitals identified in any period as best practice are more likely to be those which, by chance, had low disutility rates or cost per admission.

In the two-stage funding mechanism proposed in chapter 6, a second stage payment acts as a buffer between funding with observed best practice at a scheduled value for quality (avoiding disutility events), and case-mix funding. This buffer was suggested, in part, to allow for sampling variability, and, in part, as a potential stopping mechanism for increasing the value of quality in the schedule to feasibly remaining within current case-mix funding[2] per admission.

Ideally, a minimum level for this buffer payment per admission would reflect the difference between identified best practice and underlying best practice. However, in determining the minimum feasible industry level for this second stage payment, the effect on the frontier of sampling variation ideally needs to be considered to allow underlying, as opposed to observed, best practice to be estimated.

If underlying, as distinct from observed (sampled), best practice can be estimated, then potential is also created for policy makers to explore questions of the absolute level of funding required at a net benefit maximising level of quality of care. Answering this question of a minimum funding level for feasible net benefit maximising care for each inpatient clinical activity (DRG) would, in turn, also allow a

[2] Case-mix funding is implicitly based on expected average cost across hospitals ignoring quality of care, as outlined in section 1.4.

more robust consideration of allocative efficiency questions in the distribution of resources across activities, with intrinsically different levels of uncertainty in outcomes.

Research to allow for sampling variability in estimating the position of the underlying frontier of best practice, therefore, has potential value to policy makers in assessing performance within, but also across, activities, in attempting to maximise net benefit.

In using data envelopment analysis (DEA) the effect of sample variation on the frontier of best practice can:
(1) attempt to be minimised;
(2) be conservatively adjusted for prior to performance measurement, with methods such as Bayesian shrinkage estimation;

Alternatively, stochastic frontier analysis (SFA) can be used in attempting to parametrically allow for sampling variability. Each of these potential approaches, in allowing for uncertainty and their advantages and limitations in estimating the best practice frontier, are now considered.

9.3.1 Minimising the effect of uncertainty on the DEA frontier

Strategies which can be used in attempting to minimise the effect of sample variation in the non-parametric DEA option include:

1. restricting frontier peers to hospitals with a minimum number of admissions and;
2. pooling data for hospitals over time to estimate the frontier technology.

However, while these strategies can help minimise the effect of sample variation on the frontier, they both result in other limitations being imposed on the analysis. In the first strategy, the comparison set is reduced, and a method for determining a minimum number of admissions needs to be established *a priori*, to prevent potential for selection bias in discretionary selection of comparator sets. In the second strategy, in any pooling across time periods, the frontier will be influenced by any technology change and price changes over time.

Therefore, while these strategies can reduce the influence of uncertainty, there is a trade-off in each case and sampling bias cannot be eliminated in estimating the position of the frontier. Therefore, approaches to explicitly allow for uncertainty with stochastic frontier analysis, or Bayesian shrinkage estimation, may be considered more fruitful.

9.3.2 Allowing for sampling variation with Bayesian shrinkage estimation

Bayesian shrinkage estimation methods can explicitly allow for sampling variation in adjusting disutility event rates across hospitals at a clinical activity (DRG) level, using hierarchical (random effects) models with maximum likelihood methods, as described and illustrated by Coory and Gibberd (1998) and Howly and Gibberd (2003).

In general terms, shrinkage estimation methods allow for sampling variability across hospitals by taking each individual hospital's disutility event rate and using the summary results for all hospitals to obtain an improved estimate of each hospital's underlying rate. Each hospital's observed disutility event rate moves towards an underlying mean rate. Shrinkage of variation towards the underlying mean is greatest for small hospitals, which have greatest intrinsic variability. Advantages in applying Bayesian shrinkage estimation to disutility events rates used as quality indicators are suggested by Simpson, Evan, Gibberd, Heuchean and Henderson-Smart (2003:257) to include:

(1) minimising the mean square error of the parameter estimate across all units (Efron and Morris, 1975);
(2) accounting for regression to the mean of individual units (Christansen and Morris, 1997) and;
(3) taking account of variation in sample size (Armitage and Berry, 1994:149-153).

Disutility event rates adjusted for sampling variation with shrinkage could, therefore, be used in allowing an underlying frontier to be estimated more conservatively in identifying best practice.

An assumption made in using shrinkage estimation methods is that the observed rate for each hospital around the 'true', or underlying, rate follows a parametric

distribution[3]. An assumption is also made that each hospital is drawn from a population of such hospitals, allowing the variance of the distribution to be estimated from observed behaviour with maximum likelihood methods. In applying shrinkage to disutility event rates in the cost disutility plane, the assumption of exchangeability of rates between hospitals requires prior adjustment for differences in case-mix and effects beyond-separation. Ideally considering effects of sample variation would allow for the bivariate distribution between costs and disutility events per admission. Stochastic frontier analysis in allowing for such bivariate distributions may therefore be preferred.

9.3.3 Modelling uncertainty using stochastic frontier analysis

In estimating a frontier minimising cost and disutility events per admission, specified under the correspondence theorem, stochastic frontier analysis (Aigner, Lovell and Schmidt, 1977; Museen and van den Broek, 1977; Jondrow, Lovell, Materov and Schmidt, 1982) could be considered in parametrically modelling uncertainty. Applying SFA methods, where there is a single output and multiple inputs, a frontier and technical inefficiency in cost disutility space can be estimated in the general form of equation (8.1) for decision making units ($i=1,..,n$):

$$y_i = f(x_i; b)\exp(v_i + u_i) \qquad (8.1)$$

where:

y_i is the output for decision making unit *i;*

$f(x_i; b)$ represents the functional form of technology, with b a vector of coefficients to be estimated and;

technical inefficiency (u_i) is assumed independent of the independently and identically distributed random variable (v_i).

SFA requires specifying a functional form for the frontier $f(x_i; b)$ and a distribution for technical inefficiency as a non-negative (one sided) random variable (u_i) as distinct from error (v_i). While providing the potential to overcome sample variation, the trade-off is the potential for mis-specification of the functional form of the

[3] The beta binomial distribution is suggested by Howly and Gibberd (2003:326) as preferable to the Gamma-Poisson distribution (Coory and Gibberd, 1998) on theoretical and empirical grounds.

technology in estimating the frontiers shape, and the functional form of inefficiency as distinct from that of chance in estimating the position of the frontier (Green, 1993; Färe, Grosskopf and Lovell, 1994). In considering the distributions of the two error components of inefficiency and noise, *a priori* justification or powerful statistical tests do not always exist to distinguish selection between alternative one-sided distributions for inefficiency (Coelli, 1995). What is attributed to inefficiency and what to chance, may, therefore, largely be determined by non-testable assumptions, with respect to the distribution of inefficiency.

In applying SFA methods to estimate a frontier, allowing for quality of care, with disutility events specified as an input, equation 8.1 would take the form in (8.2) below, with output of admissions for hospital i, y_i as a function of cost[4] (c_i) and disutility events (vector du_i):

$$y_i = f(c_i, du_i; b)\exp(v_i + u_i) \qquad (8.2)$$

In interpreting the production function in (8.2), disutility event rates represent quality of care under the correspondence theorem's conditions. Lower disutility event rates reflect higher quality of care. Therefore, hospitals with higher resource use, or cost, per admission should be able to produce lower disutility event rates.

Equivalently, hospitals with lower quality of care (reflected in higher disutility event rates) may be able to reduce costs per admission. However, at some point, lowering quality of care can lead to both disutility events rates and cost of care per admission increasing, as considered with congestion efficiency modelled under DEA, in section 5.6.3. However, if disutility event rates are determined by quality of care, then hospitals with twin objectives of cost minimisation and health maximisation should not be producing in this region.

An assumption implicit in equation (8.2) is that disutility events are determined by quality of hospital care. Clearly this requires that disutility event rates are adjusted for differences in risk factors across patient populations treated, and effect beyond-separation in satisfying the correspondence theorem. As discussed in chapter 3, at a

clinical activity level patient populations have greater homogeneity and adjustment for differences in risk at admission and effects post-separation can be undertaken flexibly and comprehensively. Case-mix adjustment can focus on adjusting for differences in expected disutility event rates and costs of care for within DRG patient risk factors, rather than differences between different clinical activities.

DEA with weak and strong disposability of disutility events, specified as inputs, allows estimation of congestion efficiency, if disutility event rates reflects endogenous choice of quality this can be interpreted as allocative inefficiency (as described in 5.6.7), rather than differences in environment.

Unlike DEA, standard SFA does not currently allow estimation of technical efficiency under weak disposability. Using SFA, may, therefore, be seen as problematic as hospitals that operate at a level of quality below cost minimisation (above a cost minimising level of disutility event rate) can influence the shape and position of the frontier. The influence these firms have on the frontier is appropriate if it reflects sampling variability, but not if it reflects choice of quality of care. In comparison, using DEA, firms in these regions can only influence the position of the frontier under the assumption of weak disposability. In general, the ability with DEA to separately estimate congestion inefficiency, and not include it as technical inefficiency, can be seen as an advantage over SFA in allowing for quality of care represented by disutility events.

9.3.4 Empirical consideration in applying SFA

Empirically, in reviewing the use of SFA to estimate efficiency of UK hospitals at an aggregated level, Street (2003) found that relative performance was not robust to choices of error distribution, functional form or model specification. Similar theoretical and empirical concerns have also been raised by Folland and Hofler (2001) in application of SFA methods in hospitals, in relation to the choice of functional form, and structural differences between cost functions of different hospital types.

[4] Ideally, physical inputs and their prices if they were available, as discussed in section 5.5.4.

With respect to allowing for quality of care in performance measurement, Newhouse (1994:318-319) suggested that the SFA approach of Zuckermann, Hadley and Lezzoni (1994), adjusting costs of hospitals with mortality rates in the upper and lower deciles, could allow for quality effects represented by differences in case-mix adjusted mortality rate. However, this approach does not stand up to critical examination.

Adjusting expected costs for hospitals above, and below, threshold levels of event rates does not recognise a trade-off between cost and value of quality of care. Disutility event rates are treated as though they were determined environmentally rather than by the quality of care. Generally, in measuring hospital performance, disutility event rates do not allow quality of care in measuring hospital performance. The lack of appropriate incentives this creates is highlighted in the study of Zuckermann et al. (1994), where coefficients for variables representing high and low mortality hospitals shared the same sign. Hospitals have their expected costs adjusted upwards, relative to other hospitals, in allowing for mortality differences when they have either low or high case-mix adjusted mortality rates.

If used to pay hospitals, this would suggest paying more to low, as well as high, quality care, given very low quality leads to higher costs in the observed hospitals. For hospitals with case-mix adjusted mortality rates just below the upper decile, perverse economic incentives would hence be created to increase mortality rates to have measured performance improved and receive additional payments. Adjusting payments upwards in hospital in the lower decile of case-mix adjusted mortality rates (higher quality) end of the scale while creating incentives for higher quality of care may also be seen as problematic in creating highly localised incentives, which may be for quality of care above a net benefit maximising level. Therefore, while the approach of Zuckermann et al. (1994) adjusts for expected costs at high and low levels of case-mix adjusted mortality rate, the adjustment does not provide appropriate incentives for quality of care in performance measurement or funding.

9.3.5 Uncertainty in funding hospitals and risk sharing

While considerable uncertainty in disutility event rates, and hence funding, may exist for an individual clinical activity for a hospital, the role of chance diminishes in

aggregating across inpatient activities (DRGs) at a hospital level. The role of chance is further diminished over time, with hospitals moving closer to an underlying disutility event rate given regression to the mean (Freedman 1980).

However, risk sharing arrangements could be considered in payment mechanisms, particularly for small hospitals, to reduce the role of chance, while broadly maintaining incentives. While some free-riding incentives may be created by such risk-sharing, natural peer review mechanisms are inherent in risk sharing arrangements at a clinical activity level. Sharing of information and identification of best practice techniques for similar small hospitals, could even potentially aid quality gains with a co-operative culture of trust (Davies, Nutley and Mannon, 2000).

In comparing individual hospital clinical performance based on disutility event rates (mortality, morbidity), Bayesian shrinkage estimation methods have been used with hierarchical modelling at a DRG level in identifying high disutility event rate (low quality) outliers, unlikely to be explained by chance (Coory and Gibberd, 1998). Once identified, outliers are investigated to allow help to be focused in improving quality where it is most needed, assuming high disutility event rates represent incompetency in providing quality of care.

Under an activity based payment and performance measurement system such an identification of outliers is necessary, particularly as low quality of care could also be the result of hiding technical inefficiency. Targeting of high disutility event rates providers for quality improvements is also supported, as high disutility event rates will likely represent quality of care with expected health losses. As discussed in section 5.7.3, under loss aversion in prospect theory (Kahneman and Tversky, 1979), these losses have greater value than equivalent health gains, with Willan, O'Brien and Leyva (2001) finding a 2-3 fold greater value for losses than equivalent health gains.

While outliers can be identified, as Gibberd and Howley (2003) suggest, quality gains at an industry level are likely to be significantly higher in improving industry practice, than identifying outlier behaviour, which, by nature, only affects a small number of providers. In comparison, in application of the correspondence theorem to performance measurement and funding under correspondence conditions, incentives

for net benefit maximising quality of care provides disincentives for too high quality (and expensive), or too low quality care.

In allowing for uncertainty, while methods such as Bayesian shrinkage estimation and SFA may be useful in estimating the underlying position of a frontier allowing for sampling variation, their use in adjusting individual hospital rates of disutility events in funding is problematic because it creates inappropriate economic incentives.

In determining payments conditional on quality of care to create appropriate economic incentives, observed rates at a clinical activity level should be adjusted for case-mix and effects beyond-separation, where appropriate. However, sampling variability should not be adjusted for, whether with Bayesian shrinkage estimation, or otherwise, as net benefit maximising incentives cannot then be created. While uncertainty from random effects outside the control of hospital can influence hospital costs for individual clinical activities, these effects should balance out aggregating across activities and time. For smaller hospitals, risk sharing arrangements could, however, be considered.

In general, only hospitals that have systematically hidden inefficiency behind reduced quality, cost-shifting and cream-skimming should fear funding conditional on quality of care, under correspondence conditions. Where correspondence conditions are satisfied, with adjustment for patient risk factors and effects beyond-separation, hospitals are rewarded for net benefit maximisation at a scheduled value rather than cost shifting or cream-skimming. Only where predictive factors within DRG are identified by hospitals that are not controlled for, can hospitals engage in cream skimming in selection of patients within a DRG. However, if factors are identifiable then they should be able to be controlled for. At a clinical activity (DRG) level identification of risk factors should be able to be flexibly and comprehensively undertaken.

9.4 Applying the correspondence theorem in other settings

9.4.1 Other hospital inpatient activities

In chapters 5 and 6, application of the correspondence theorem to hospital performance measurement, and funding, were identified as allowing quality of care to be included at a clinical activity (DRG) level, consistent with net benefit maximisation. This was illustrated for a single hospital inpatient activity (DRG E62a), where relative quality of care was represented by mortality rates.

For other inpatient activities (DRGs), disutility events, such as post surgical complications, adverse drug reaction, readmission or linkage to post care events, may be more appropriate indicators of quality of care. The linear nature of the correspondence theorem framework allows inclusion of multiple effects, in the proposed performance measurement or funding, consistent with net benefit maximisation under correspondence conditions.

As described in section 5.7, to allow a comprehensive and systematic approach, decision analytic methods, as applied in health technology assessment (Weinstein and Fineberg, 1980; Hunink, Glasziou, Seigel, Weeks, Pilskin, Elstein and Weinstein, 2001), can be employed, to identify the effects of care. These effects then require measurement, using data linkage, where necessary.

9.4.2 Application in other health settings

Use of the correspondence theorem in measuring relative performance, and funding conditional on quality of care, has been described and illustrated at a clinical activity level for inpatient care in hospitals. The generalized nature of the correspondence suggests analogous application to other health care settings, where quality of care can be measured by reduction in disutility events. As in hospitals, quality of care indicators should reflect the objectives of care and be able to be framed as verifiable disutility events.

In sub-acute, and non-acute, health care settings, the ability to reframe utility bearing indicators of quality as disutility events increases in importance. In settings such as nursing homes and rehabilitative care, measures of quality of care in the broader sense

of the WHO (1948) definition of health, including physical, mental and social well-being, are likely to represent better indicators of quality than mortality or morbidity.

In allowing inclusion of utility aspects of the proportion of patients not meeting a standard, level of incapacity in functioning or direct disutility can be used as disutility event rates, as was outlined in section 5.7.2.1. Adjustment for differences in patient risk at admission can then either be undertaken on the rate not meeting a standard, or the standard itself.

If a finer distinction is allowed with measurement of health related utility or functioning on a cardinal scale with defined endpoints, then disutility events can be framed as limitation in degree of functioning or disutility respectively.
Possibilities for disutility event 'rates' then include:

(1) average functional limitation, or disutility at separation, standardised for risk at admission and potentially baseline predictive factors (as indicators of capacity to benefit) and;

(2) one less average utility over a given time frame standardised for baseline risk.

Framing health effects as disutility event rates, to include as inputs, allows more meaningful performance measurement than including health effects as utility bearing outputs. Applying the correspondence theorem, an input specification provides an underlying objective function of net benefit maximisation, rather than, at best, an implicit objective of average cost-effectiveness with an output specification.[5] As described in section 5.7.2.3, an underlying objective function of average cost effectiveness is problematic both because of floor effects from health as a stock variable, which care incrementally effects, and the non-tradable nature of health effects in a patient population.[6] Maximising net benefit allows for both these characteristics in reflecting incremental cost relative to incremental effect, at the decision maker's monetary value for effects of care.

Integration of effects of care over time may particularly be important in health care settings such as nursing home care, and other sub-acute health care setting activities,

[5] Cost effectiveness in the case of a single disutility event

to the extent that they involve palliative care. Measuring relative effects in such forms of care should reflect objectives of care, with, for example, death not necessarily seen as a disutility event in palliative care and process aspects as important. Again this highlights the importance of a disaggregated (clinical activity) level of performance measurement, in allowing for differences in objectives and, hence, appropriate measures of the effects of care.

Ideally, process aspects of care, such as self respect or dignity during care should also be included. As section 5.7.2.2 described, where process aspects can be framed as not meeting a standard, disutility events or continuous measures such as waiting time, they can be included in addition to health effects of care implicit in net benefit maximisation. These process aspects can then be valued using contingent valuation methods, as described in section 5.7.2.7. .

An example of the potential use of functional limitation as a cardinal measure of health in practice is provided by use of the functional independence measure of Lee, Eagar and Smith (1998). This measure is currently employed at admission and separation in sub-acute, and non-acute, settings in New South Wales. Questions that need to be addressed by such instruments in satisfying correspondence assumptions creating appropriate incentives for quality of care include:

(1) whether scores reflect effects of care across populations in a cardinal manner;
(2) the ability to verify measures and hence scope for finessing by providers and;
(3) whether differences in patient capacity to benefit are able to be adjusted for.

In adjusting for patient capacity to benefit, Roos (2002) identified an approach using a Malmquist output quantity index to measure changes in state of health (as physical or anatomical condition of the body) from presentation at admission to separation, with the patient as the producer of health. Hospital services were modelled as inputs in production of multiple 'functionings' by patients affecting living conditions (or their

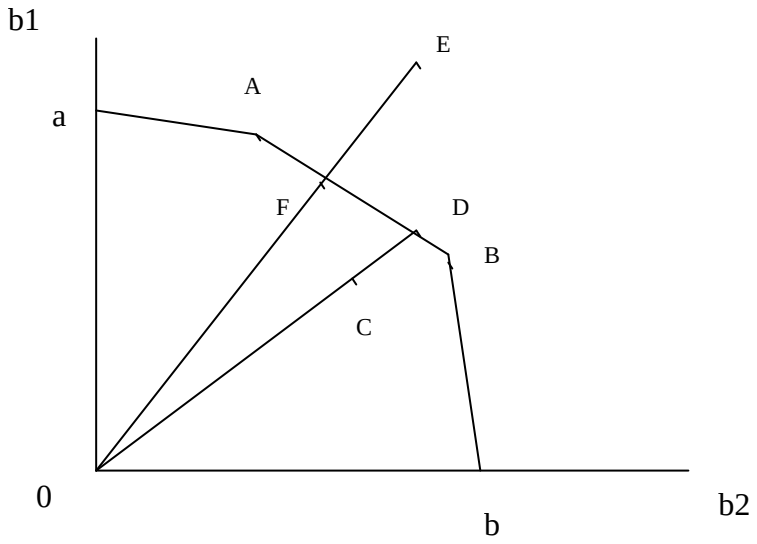
[6] As discussed in detail in section 5.7.2.2.

restriction). Following Sen (1985, 1993), functionings were defined as reflecting the achievement of a person in the way she manages "to do", or "to be", given commodities possessed and ability (capabilities) to transform these into functionings.

Roos (2002) constructed a frontier of potential functionings, given capability prior to treatment, across individuals undertaking the same treatment. Changes in any individuals functioning between admission and separation were then measured as a ratio of distance functions relative to a frontier of capabilities at point of admission. In figure 9.1 (taken from Roos (2002:257)), an individual at point C before treatment, and point E post treatment, would have an output index I given by (9.1):

$$I = \frac{0E / 0F}{0C / 0D}. \qquad (9.1)$$

Figure 9.1 Illustration of activities before surgery and changes in activity due to treatment (Roos, 2002:257)



Where:

b1 and b2 are activities, for example visual acuity and contrast sensitivity in the case of cataract surgery, and;

the frontier aABb and the set of points to the southwest represent possibilities before treatment across a patient population.

The Malmquist output index, in equation 9.1, provides a relative measure of output for each admission (patient) across patients treated, allowing for multiple positive (utility bearing) aspects of functioning. To allow for predictive differences in patient capability, modelling of separate frontiers by predictive characteristics, such as age and sex, were suggested.

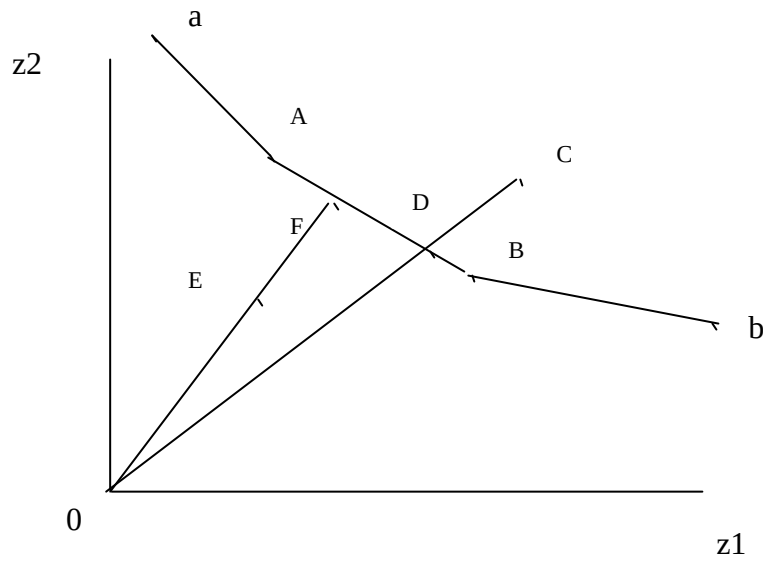
Roos also considered adaptation to allow for bad activities, noting that:

> "In many cases the result of medical treatment is in terms of a reduction in bad aspects of performing activities". Roos (2002:257)

While Roos (2002) considered bad aspects, alongside good aspects, as outputs, multiple disutility events representing functional limitations could also be modelled as

'bads' in a Malmquist input index. Figure 9.2 illustrates a frontier from which such an index could be constructed for an individual at C pre-treatment and E post-treatment.

Figure 9.2 Limitations in activities before surgery and changes in activity limitation due to treatment



Where: z1 and z2 are rates of ***limitations*** in activities and;

frontier aABb and the set of points to the northeast represent possibilities before treatment across a patient population.

A Malmquist input index, representing relative change in inputs measuring limitation across individuals given their capabilities could then be given by (9.2):

$$I = \frac{0E / 0F}{0C / 0D} \qquad (9.2)$$

The index in equation 9.2 offers one method for allowing for capacity to benefit with a non-negative change in health limitation as an input. Specifying health effects as disutility effects (limitation) as an input, allows comparison of incremental, rather than average, cost effectiveness in relative performance measurement under a utility bearing output specification of health effects. To allow for capacity to benefit in including quality of care effects over time, further research is suggested in adapting

the general approach of Roos (2002), to allow relative change in health limitation as an input.

9.4.3 Applying the correspondence framework in other service industries

As a general theorem, the correspondence theorem can be applied to performance measurement and funding mechanisms in service industries, where effect of quality is measurable as disutility bearing events avoided. The economic objective of maximising net benefit per service, underlying the correspondence theorem, provides a framework for comparing the value of incremental effects relative to incremental costs of quality of care. In service industries, such as health care, where there are incremental non-tradable effects of services, this economic objective is more appropriate than (minimising) average cost effectiveness in assessing relative performance, allowing for quality. In public sector services, prices are not available to represent value of quality differentiated services. As was argued in chapter one, in the case of hospital care, the transactions conditions under which prices of services could reflect quality of care are also not present.

Application of the correspondence theorem is likely to be particularly appropriate in services industries where quality is naturally represented by reduction in disutility events. In public service settings, these include industries such as job placement agencies, where disutility events could include lack of placement, mismatch or turnover of placements, and corrective services where re-offence or recidivism rates reflect quality of service. While the correspondence theorem provides a method for valuing quality of care consistent with maximising net benefit in such service industries, it also provides a framework for explicitly addressing prior risk and post-separation effects in preventing incentives for cream-skimming and cost-shifting.

In private industry, prices for services do exist and hence revenue and profit efficiency could be calculated. However, the extent to which prices reflect service quality will largely depend on transaction conditions of the industry in question. In many service industries, transaction conditions consumers' face may not allow them

to distinguish between differences in quality of services ex-ante, but also ex-post[7]. Transaction conditions of information asymmetry, complexity of decision making, small numbers problems with customisation of services and bounded rationality (1957) present in a hospital setting may be characteristic, while less pronounced in other service industries.

Net benefit maximisation may also be considered a more appropriate measure of relative performance than maximising profit efficiency in industries, where prices for quality differentiated outputs exist, but do not necessarily reflect effects of services due to transaction conditions.

While demand and prices can reflect consumers' preferences for quality, this largely depends on the extent to which transactions conditions are present for consumers to distinguish the quality of provider services. Where transactions conditions are such that prices do not necessarily reflect differences in the quality of services or goods and the approach provides a framework for consistent valuing of measured effects of quality. This can in particular be the case where asymmetry of information exists, as in Akerlof's 'Lemons' (Akerlof, 1970).

The application of the correspondence theorem to performance measurement, and in the case of publicly provided services, funding, are, therefore, suggested to be potentially fruitful in industries where:

(1) quality can be represented by rates of disutility bearing effects from services;

(2) the effects of quality are incremental and;

(3) lack of information and other transaction conditions facing consumers does not allow quality to be reflected in demand or prices.

In comparison of relative hospital performance, net benefit maximisation is clearly a more appropriate objective function than average cost effectiveness, given incremental and non-traded effects of quality of care. Even if there were prices for

[7] This inability to identify quality ex-post as well as ex-ante was argued in Chapter One for individuals receiving hospital care, given effects of care under uncertainty, small numbers problems with

each hospital, for each output, the derived nature of demand and transaction conditions characterised by a-symmetry of information with respect to quality of care suggest advantages of net benefit maximisation relative to profit maximisation. However, average cost effectiveness, or profit maximisation, may be appropriate in other industries where such characteristics or transaction conditions are absent.

Service industries, where such transaction conditions may prevent or distort service quality (measurable by disutility event rates) being reflected in prices and demand, include:

information technology systems and services (failure rates);
security device and service industries (detection rates);
repair and maintenance (detection and failure rates) and;
production line rejection (non-compliance or undetected defect rates).

In each case, these industries have a cost per service which can be traded off against effects of quality, measured by disutility events over time. The correspondence theorem allows such a trade-off and a linear framework in which baseline risk and effects, including those post-separation, can be simply and intuitively incorporated. To the extent that effects of such services have floor effect and are non-tradable, the implicit objective of net benefit maximisation underlying the correspondence theorem is more appropriate than average cost effectiveness in reflecting relative performance.

9.5 Summary

In chapter 9, future research in extending and improving application of the correspondence theorem and its use in other settings, has been critically considered.

Robust application of the correspondence explicitly requires satisfaction of correspondence conditions, with adjustment for patient risks and coverage by disutility events of effects of care. A four stage process to systematically enable this at a clinical activity level was identified. Using decision analytic methods at a clinical activity level was suggested to allow comprehensive identification of effects, including those beyond-separation reflecting objectives of care. The second stage

---

idiosyncrasies of patient in determining risks and the absence of information on counterfactual

involves measuring identified effects, including data linkage where appropriate to effects beyond separation. These effects can then be reframed as disutility event rates in the third stage. Finally differences in patient risk factors at a clinical activity level can be adjusted for directly with regression or in restricting comparators, for example with use of non-discretionary variables representing expected costs and risk of disutility events in DEA. Adjusting for differences in environmental variables, scale, access to technology, or fixed nature of inputs in the short run were also considered using standard methods of peer grouping, second stage regression, VRS formulation and the use of non discretionary variables.

In considering data limitations in application, use of a clinical activity level for performance measurement requires accounting frameworks to allocate joint costs to individual activity level. In Australia, accounting frameworks have already been constructed in calculating DRG case-mix weights (NHCDC 2000). Where costs are decomposable in physical inputs and their prices, DEA frontiers could allow for price differences and estimate relationships between inputs and quality in reflecting reflect the tradeoffs between inputs and disutility event rates. The limitation of the illustrated use of the correspondence framework with available costs was shown to be no greater than that in current performance measurement, based on minimising cost per case-mix adjusted admission.

To allow for uncertainty in estimating underlying best practice with frontiers minimising cost and disutility event rates, future research, using stochastic frontiers analysis and Bayesian shrinkage estimation techniques were suggested. While adjustment for sampling variation with approaches such as Bayesian shrinkage estimation is appropriate in estimating an underlying frontier for buffer payments, divergences from net benefit maximising incentives are created if they are used to determine first stage payment, and particularly for smaller hospitals. In allowing for uncertainty in paying individual hospitals options for risk sharing arrangements between hospitals with smaller numbers of admissions might also be considered. However, in general, provided correspondence conditions are satisfied, with adjustment for baseline risks and event shifting, only hospitals which have

---

alternative service provision.

systematically hidden inefficiency behind reduced quality, cost-shifting and cream-skimming should fear funding, as proposed in chapter 6, conditional on quality of care.

While the correspondence theorem has been demonstrated in hospitals using frontier methods in relative performance measurement with quality of care presented by disutility event rates, future research is suggested in applying the correspondence theorem in other settings where net benefit maximisation per service is an appropriate objective. In service industries, where quality of service can be represented by disutility event rates, the correspondence theorem can be applied to performance measurement and funding consistent with maximising net benefit per service. This may be particularly useful in industries where effects of care are incremental[8], and prices either do not exist, or reflect quality of service, due to the transaction conditions that consumers face.[9] In the public sector, application to performance measurement and funding is suggested in other health care settings and industries such as corrective services (recidivism) and employment placement services (unemployment or job retention). In each of these industries, as in hospitals, conditions of a common comparator and coverage of effects create an explicit framework for adjusting for baseline risks of customers and effects of services, over time, to prevent incentives for cream-skimming and cost-shifting. In the private sector, directly including effects of services in representing quality may be valuable in industries where transaction conditions preclude quality being reflected in prices of services. These include service industries such as maintenance and safety (failure rates); information technology services (downtime) and security devices and services (detection rates). The ability to reframe utility bearing aspects of care, where measurable as standards of service, utility or functional ability, suggests potential for wider application still.

[8] In relative performance measurement including quality of care, cost effectiveness is not meaningful where effects of services are incremental as discussed in section 5.7.2.3.
[9] Profit maximisation is not possible if prices do not exist for market outputs and may not be desirable if consumer value is not reflected through prices if they were available.

Chapter 10: Conclusion

Problems of perverse economic incentives for lowering quality of care, and cost-shifting beyond point of separation created in funding and measuring public hospital performance with activty based measures, such as cost per case-mix adjusted separation are well recognised. In Eckermann (1994) a research agenda for addressing these problems was proposed. This agenda suggested the need for performance measurement to reflect the twin objectives of minimising costs and maximising health gain following Harris (1977), whilst allowing for effects within a health care system, given incomplete vertical integration of patient care (Evans, 1981).

Motivated by this research agenda, the primary economic question addressed in this thesis has been how to include quality of care, represented by disutility-bearing event rates such as mortality, morbidity and readmission, in comparing economic performance and funding of public hospitals in providing inpatient care. In addressing this economic question, the policy objective has been to identify performance measures and payment mechanisms that provide appropriate economic incentives for quality of care. In identifying appropriate and robust measures, questions have also been addressed in relation to level of aggregation, adjustment for patient risk factors (in preventing incentives for cream-skimming) and allowing for effects of care beyond-separation to avoid incentives for cost-shifting.

Currently, where funding is based on case-mix payments, and relative hospital performance measured with cost per case-mix adjusted separation at a hospital level, the costs of quality of care within-admission are included, but their effects are ignored. This failure to hold hospitals economically accountable for quality of care creates economic incentives for cost-minimising quality of care for each clinical activity, implicitly valuing health effects of quality at 0.

For any given clinical activity, economic incentives for quality of care minimising cost per admission may be mitigated to differing degrees, by regulation of clinical practice such as processes of accreditation and clinical peer review, with associated partial clinical

performance measurement. For any given clinical activity in any hospital, depending on medical service provider (clinician), health arguments related to their agency role under the Hippocratic Oath and professional standing in their objective function, can also play a role in opposing cost minimising quality of care in practice, depending on the relationship with hospital administrators attempting to minimise costs per admission.

However, as the mechanisms which mitigate against economic incentives for cost minimising quality are dependant on local circumstances and individual's values in any hospital for any given clinical activity, and hence discretionary, scope is created to skimp on quality. By accounting for the costs, but not the effects of quality, a moral hazard is created to hide technical inefficiency behind lower quality of care. For any given clinical activity (DRG), cost minimising quality of care does not require cost minimisation, where case-mix funding is based on average within-admission cost for each inpatient activity (DRG) and average industry quality of care is above that of cost minimisation.

From a health system perspective, incentives for hospitals to provide quality of care minimising costs per admission, for any given clinical activity, are particularly problematic considering effects beyond-separation. Reduction in quality to minimise cost per admission may not have a clearly defined endpoint, where there is discretion with respect to point of separation of patients and a lack of accountability for effects beyond-separation.

Problems of inappropriate economic incentives for quality of care and cost-shifting with activity-based funding, such as case-mix funding, and measures ignoring health effects of care such as cost per case-mix adjusted separation, are not new. Harris (1977), Evans (1981), McGuire, Henderson and Mooney (1988) and Eckermann (1994) have expressed similar concerns. While this thesis has shed some additional light on the nature of these problems, this has been incidental to identifying and illustrating an alternative framework and method for relative performance measurement and funding, which systematically addresses these problems.

In addressing the question of how disutility events can be included in performance measurement, to create appropriate incentives for quality of care, the main findings in this thesis are as follows:

1. In trading off the costs and value of quality of care, minimising cost per admission conditional on quality of care is a necessary condition for economically efficient (net benefit maximising) provision of health care, but minimising cost per admission *per se* is not. This was illustrated theoretically in chapter 2, and empirically in chapters 5 and 6. The only circumstance in which cost minimising quality of care allows an appropriate trade-off (corresponds with net benefit maximisation), is where infra-marginal cost reductions, from better outcomes within-admission, continually outweigh the intra-marginal costs of providing higher quality of care. Where there are diminishing marginal returns to increasing quality of care, or discretion of providers with respect to point of separation, quality of care minimising cost per admission will not correspond to that of net benefit maximisation in practice. To create economic incentives for appropriate quality of care, performance measurement and funding need to hold hospitals accountable for effects of care (such as mortality, morbidity and readmission), under an appropriate trade-off with resource use, or costs of care.
2. In measuring relative performance of hospital inpatient services, allowing for quality of care, a lower level than the hospital is required to identify and attribute effects from quality of care in heterogeneous clinical activities, given more than 660 diagnostic related group (DRG) inpatient services. At a clinical activity level, decision analytic methods can be used to allow a flexible, comprehensive and evidence-based approach to including health effects of quality of care. A clinical activity level, therefore, provides the ability to overcome a performance-efficiency paradox, where in ignoring quality of care, lowest cost per admission is measured as efficient, but may reflect quality of care minimising cost per admission and represent worst performance (lowest health gain and highest costs), in treating patient populations over time. Identifying relative performance, peers and benchmarks at a clinical activity level also allows a problematic aggregation

paradox (Fox, 1999) of aggregate economic efficiency measures, such as cost per case-mix adjusted separation, to be overcome, as was described in chapter 3 Inefficiency hidden by aggregation is also revealed and the use of problematic case-mix cost weights to aggregate admissions is avoided.

3. While a clinical activity level allows flexible identification of effects (such as mortality, morbidity, iatrogenic events, readmission), the question of how to specify these disutility events in performance measurement needs to be addressed. Output specifications of disutility events were illustrated, in chapter 4, to not allow appropriate incentives for performance measurement in representing quality of care. Use of inverted rates of disutility events, such as inverted readmission rates, as used by Bruce and Gregan (1997), while demonstrated as able to be normalised by admission to avoid size biasing, create dichotomous incentives for high and low quality of care. These incentives are inherent in the monotonically increasing value of a marginal health gain under this specification of disutility events. An alternative specification, of admissions without disutility events (for example survivors), at best (for a single disutility event) represents an underlying economic objective of minimising average cost-effectiveness. However, an economic objective of minimising average cost-effectiveness is problematic in comparing relative performance of hospitals, in not allowing for the stock nature of health, and incremental and non-tradable nature of health effects in patient populations treated. Lastly, the hyperbolic method of Färe, Grosskopf, Lovell and Parsuka (1989), equi-proportionally contracting weakly disposable undesirable outputs (e.g. pollution) and expanding marketed strongly disposable desirable outputs (e.g. electricity generation), was considered. This approach was suggested to not be meaningfully adapted to where disutility event rates represent effects of quality of care. In the case of hospitals, assuming desirability of admissions *per se*, denies the derived nature of demand for care, given the intermediate nature of hospital care in affecting health. In interpreting technical efficiency as a performance measure, regions of the hyperbolic frontier where admissions decrease and disutility event rates increase, are also problematic, unless disutility event rates are assumed to be externally determined. This

becomes particularly problematic, as output-orientated economic efficiency is unable to be estimated with the hyperbolic method and hence technical efficiency becomes the de-facto performance measure. Improvement in technical efficiency, where disutility event rates and costs per admission increase, cannot be subsumed into allocative inefficiency. In the absence of a market price for public hospital admissions as a 'desirable output' a monetary shadow price of avoiding disutility events, is also not estimable using the method of Färe, Grosskopf, Lovell and Yaisawarang (1993).

4. The incremental and non-tradable nature of health effects of care in patient populations treated, suggests, as in health technology assessment (Drummond et al. 1987, Drummond et al. 1997), that incremental cost effectiveness, relative to a threshold value, is the appropriate relative performance measure for hospitals. An economic objective of maximising net benefit (Stinnett and Mullahy, 1998) per admission, as the incremental value less incremental cost per admission, implicitly allows this appropriate trade-off between relative health effects and costs of providers. Net benefit maximisation reflects the same preferences in ordering hospital performance as those implicit in assessing incremental cost effectiveness ratios, relative to a decision maker's monetary value of health effects of care.

5. A direct specification of net benefit maximisation, as an efficiency measure, is not possible, given the requirement for non-negative inputs and outputs in ratio measurement. However, a linear transformation allows a relative performance measurement corresponding to maximising net benefit per admission.

   Maximising net benefit per admission has a one-to-one correspondence with minimising per admission costs plus disutility events valued as in net benefit (at the decision makers' willingness to pay threshold), where:

   (i) Compared hospitals face a common comparator, or equivalently, differences between hospitals in expected costs and disutility event rates are adjusted for, and;

   (ii) Disutility events cover effects of care.

6. The correspondence theorem allows a trade-off between health maximisation and cost minimisation, consistent with maximising net benefit per admission under correspondence conditions. As illustrated in chapter 5, using frontier methods at a clinical activity (DRG) level, under a specification with inputs of disutility events and costs (ideally resources) and an output of admissions, policy makers are able to be informed of:

    (i) relative economic performance and peers conditional on decision makers value of avoiding disutility events (quality of care), including a 0 value representing minimising cost per admission;

    (ii) technical efficiency of net benefit relative to a frontier minimising cost and disutility events per admission;

    (iii) regions of willingness to pay for disutility events over which hospitals on this frontier have best practice;

    (iv) industry shadow price of avoiding disutility events;

    (v) decomposition of technical efficiency, under constant returns to scale into technical efficiency under variable returns to scale and scale efficiency and;

    (vi) decomposition of technical efficiency, under strong disposability of all inputs and outputs, into technical efficiency where disutility events are assumed weakly disposable and congestion efficiency[1].

    For a known decision maker's value of willingness to pay, the method was also demonstrated to allow:

    (i) a decomposition of economic efficiency, consistent with net benefit maximisation, into technical efficiency of net benefit in minimising cost and disutility events, and allocative efficiency in valuing effects relative to costs of quality of care (disutility event rates), and;

    (ii) estimation of differences in net monetary and health benefit by comparing isocost curves in the cost disutility plane.

[1] As discussed in section 5.6.3 such congestion inefficiency can be seen as a form of allocative inefficiency where disutility event rates represent quality of care and are assumed endogenously determined.

7. Performance measurement under correspondence conditions can provide appropriate incentives for quality of care in benchmarking and peer identification. To create accountability and appropriate economic incentives for quality of care requires funding mechanisms to reflect an underlying objective of net benefit maximisation. Current proposals for allocating a fixed level or proportion of funds based on targets, such as the 'reward results' initiative in the USA (NHCPI 2002), can only potentially provide localized incentives, and lack an economic basis for determining this payment level. In comparison, a funding mechanism conditioned on effects of care (disutility event rates), with scheduled payments at the decision maker's value of willingness to pay, creates continuous and exact incentives for net benefit maximisation, under correspondence conditions. A two-stage sequential funding mechanism, identified and illustrated in chapter six, provides a method of moving incentives from cost minimising towards net benefit maximising quality of care, while maintaining budgetary control at current case-mix funding per admission. A first-stage payment relative to best practice, initially scheduled at the shadow price for industry value of quality, rewards quality implicit in industry behavior, while maintaining case-mix funding per admission. A second-stage payment per admission as a residual of case-mix funding (based on expected average cost) per admission allows a smooth transition from average cost based case-mix funding. This two-part sequential funding mechanism provides policy makers with a method to undertake a planned and manageable culture change, from case-mix funding to funding conditional on quality of care, that:
    (i) provides increasingly appropriate incentives in valuing quality of care (allocative efficiency);
    (ii) reduces scope to hide technical inefficiency hidden behind low quality of care in gradually making hospitals accountable for effects of care and;
    (iii) ensures budgetary control is maintained by starting at the current shadow price for avoiding disutility events and stopping increases

in value of quality when a second stage buffer payment is exhausted.

8. Satisfying correspondence theorem assumptions of a common comparator and coverage of effects of care, provides a framework to avoid incentives for cream-skimming and cost-shifting, while creating incentives consistent with net benefit maximisation. In applying the correspondence theorem, effects of care can be included as disutility events where they can be framed either as: not meeting a standard; cardinal measure of functional limitation, or; disutility directly. The linear nature of the correspondence theorem allows inclusion of multiple effects, including those beyond care, in satisfying coverage of effects of care. At a clinical activity (DRG) level, decision analytic methods can be employed in identifying prognostic factors and effects of care, both within and beyond separation, and for risk factors that can be flexibly identified and adjusted for at a patient level within-DRG. Adjustment for patient prognostic factors at admission, and data linkage to effects beyond-care, can, therefore, be comprehensively undertaken in a systematic, evidence-based manner at a DRG level. A DRG level of performance assessment can also naturally link with processes such as clinical audit, peer review, benchmarking and accreditation, in verifying measured effects of care to prevent incentives for gaming or finessing of reported quality.
9. Where correspondence assumptions are satisfied, the proposed performance measure and funding mechanism provides an underlying objective function consistent with net benefit maximisation. This creates a more appropriate environment for resource allocation decisions within hospitals, from that characterised by Harris (1977), as a complete separation between administrators, with an objective of cost minimisation under activity based funding, and clinicians, with an objective of health maximisation. In considering effects on the internal organisation of the hospital the proposed funding mechanism changes the implicit objective of administrators from cost minimisation to net benefit maximisation. Administrators need to consider opportunity cost in terms of quality effects of care that resource allocation

decisions represent, rather than acting as accountants in attempting to minimise cost within admissions under case-mix funding. Clinicians, while having the value of the effects of their care recognised, become accountable for these effects of care, relative to observed (net benefit maximising) best practice. Under this funding system, the process of negotiation between clinicians and administrators should more closely reflect program budgeting and marginal analysis, trading off costs and effects of care. A more appropriate environment is created for a working relationship between administrators and clinicians within hospitals as well as a more appropriate alignment between objectives of policy makers, as principal, and that of hospitals implicit in the funding mechanism.

10. To assess the effects of incentives, under the proposed funding and performance measurement framework, policy makers can monitor the shadow price of avoiding disutility events as well as relative technical and economic efficiency, over time. Where a decision maker's willingness to pay for effects of quality of care is known over time, changes in net benefit can be tracked using index methods, as described in section 7.4. In allowing for budget control and equity considerations, capitation payments can be employed, as currently is the practice in many health systems, to overarch payments to hospitals in regions. Across regions, this provides a means to constrain budgets in treating patients over time, and a mechanism for distribution of resources by need. Under case-mix funding, variability in quality of care results from discretionary opposition to an implicit zero value for quality, and the barrier to adoption of net benefit maximising evidence-based medicine that this creates. Capitation payments to regions overarching case-mix funding, while attempting to create equal access for equal need, are to services of variable quality, that depend on local conditions opposing cost minimising incentives in practice. However, if capitation payments to regions overarch the proposed sequential two-stage funding mechanism, then equal access for equal need can be to services where under correspondence conditions incentives for evidence-based, net benefit maximising quality of care are encouraged in practice. In aligning with health

technology assessment, regulation and clinician objectives conditions are created for more equitable, as well as appropriate, quality of care in practice.

The correspondence theorem, applied at a clinical activity level in combination with decision analytic methods, creates a framework for systematically identifying what data is required to create incentives consistent with evidence-based medicine. The linear nature of the correspondence provides the ability to intuitively and flexibly synthesize disutility events in performance measurement, and funding, to provide incentives for quality of care consistent with net benefit maximisation. The ability to reframe utility-bearing effects as disutility events, either where there is a standard, or a cardinal, measure and to adjust for differences in risks, suggests that the coverage of effects by disutility events is not a theoretical limitation. In incremental comparisons, wherever quality of care can be represented by standards of care, cardinal measures of functional ability or direct utility bearing measures, equivalent disutility event rates can be framed.

The correspondence theorem framework, while allowing flexible inclusion of effects of care consistent with net benefit maximisation at a clinical activity level framed as disutility events, is not, however, a panacea by itself. In satisfying correspondence conditions, existing agendas for policy makers and research to overcome incentives for cream-skimming and cost-shifting are focused at a clinical activity level on:

1. modelling risk factors for disutility events and costs, to enable adjustment in preventing incentives for cream-skimming (a challenge shared by capitation systems, insurers and implicitly, case-mix payments) and;
2. undertaking data linkage, using decision analytic methods, to prevent incentives for cost-shifting (a challenge also implicitly shared by case-mix funding) and event-shifting, in allowing for effects of quality beyond-separation.

Requirements to adjust for differences in expected costs and effects (including effects beyond-separation) are explicit in applying the correspondence theorem. These

adjustments should, however, be undertaken whichever method is used to incorporate effects of quality of care in evaluating hospitals' relative performance. Empirical limitations in adjustment for patient risk factors and effects beyond-care, while explicit in applying the correspondence theorem, should be addressed with any performance measurement or funding mechanism to create appropriate economic incentives.

Pointing analysts to the requirements to adjust for these differences can, therefore, be seen as an advantage, rather than a limitation, of applying the correspondence theorem to hospital performance measurement and funding. Adjustment for patient risk factors and effects beyond-care, while explicit in applying the correspondence theorem, should be addressed with any performance measurement or funding mechanism to create appropriate economic incentives.

In health technology assessment, correspondence theorem assumptions are naturally satisfied where evidence is based on randomized control trial evidence, and effects (survival, life years, QALYs) can be reframed as disutility events (deaths, life years and QALYs lost relative to most effective technology), as illustrated in chapter 8. Adjustment for differences in patient risk factors and effects beyond-separation are, however, required to satisfy these assumptions when applying the correspondence theorem to considering relative hospital performance.

In illustrating application of the correspondence theorem to relative hospital performance adjustment for differences in patient prognostic risk factors and data linkage beyond-care to satisfy correspondence conditions was beyond the scope of data I was able to access in undertaking this thesis. However, methods to identify and adjust for patient differences at a clinical activity level were extensively explored in section 5.7. In adjusting for differences in expected costs and disutility event rates, the linear nature of minimising cost plus disutility events, valued at willingness to pay, allows direct regression of the statistic itself, using available covariates. At a clinical activity level, these covariates can include factors such as age, gender, co-morbidities, prior care and diagnostic test results. This approach is consistent with that suggested in health technology assessment by Hoch,

Briggs and Willan (2002), who illustrate regression of the linear net benefit statistic on patient covariates, in adjusting for patient risk factors and undertaking subgroup analysis.

The use of a net benefit framework (Stinnett and Mullahy, 1998), implicit in applying the correspondence theorem, challenges policy makers to identify monetary values for relative effects of care, as in health technology assessment. It is important for policy makers to convey signals of the values for effects of care in the funding and performance measurement of public hospitals.

The lack of value and economic accountability for quality of care, under case-mix funding, creates reliance on mechanisms, such as targets, licensing and performance monitoring, in attempting to mitigate against economic incentives for quality of care minimising costs per admission. At best, these mechanisms may provide localised incentives to oppose cost minimising quality of care. It is more likely that they have little impact in the absence of reward structures and formal feedback procedures, and the consequent discretionary nature of compliance. At worst, the partial nature of these mechanisms (particularly if applied at an aggregate level), can create perverse incentives to comply with measured indicators at the expense of other objectives, as Smith (1995) suggested.

Clinical effects of hospital care, within- and beyond-care (with data linkage), are increasingly being measured at a clinical activity (DRG) level in health systems such as Australia's (Holman et al 1999; The Australian Council on Healthcare Standards, 2001; National Health Performance Committee, 2000; Hargreaves, 2001). The correspondence theorem can use the same data as these monitoring and accreditation activities, to allow funding and performance measurement to create incentives for evidence-based medicine in practice. Using decision analytic methods at a clinical activity level, to comprehensively identify effects within- and post-admission, provides a more focused approach than these current frameworks. In the absence of a cohesive theoretical framework, current approaches often duplicate across different levels of aggregation and types of monitoring, yet are still characteristically piecemeal in identification and

adjustment of effects. The ability of the proposed framework to tie in with processes of health technology assessment and prevent duplication of effort across levels, and over time, has the potential to lead to a lower cost of monitoring, particularly in the long-term. The incremental cost of meeting the policy challenge to robustly satisfy correspondence conditions is, therefore, likely to be marginal; however, the incremental pay-off is significant.

There are considerable benefits in replacing incentives for cream-skimming, cost-shifting and cost minimising quality of care in public hospitals with incentives to maximise net benefit per admission. Hospitals appropriately held accountable for their quality of care, under correspondence conditions, effectively become economic agents of evidence-based medicine, with economic incentives to maximise net benefit in their choice, and use, of technologies, their system practices, and in referral with data linkage.

Without the use of a systematic framework to include quality and effects of care, such as that provided by the correspondence theorem, appropriate economic incentives will not be created. Data linkage to costs beyond-care alone creates perverse incentives for referral practices in follow-up treatment, and skimping on aspects of quality of care such as mortality within-separation, or non-treatable effects beyond-separation. In general, linkage to costs alone does not value effects of care, except in prevention of costs, creating incentives for cost minimisation at a health system level, rather than net benefit maximisation.

The correspondence theorem, applied in combination with decision analytic methods and data linkage, allows the aligning of performance measurement and funding with an objective of net benefit maximisation at a health system level. In terms of the principal-agent framework identified by Smith (1995) and Goddard, Mannion and Smith (2000) this provides the funder (principal) with an attribution, performance measurement and reward system that creates incentives for hospitals (agents) to maximise net benefit in treating a patient population within the health system.

While ultimately an empirical question, valuing and holding hospitals accountable for effects of quality of care, under correspondence conditions, can be predicted to result in improved quality and outcomes of care. Employing the two-stage sequential funding mechanism, illustrated in chapter six, quality improvement for any clinical activity (DRG) can be achieved whilst maintaining an industry cost per admission equivalent to current case-mix funding. Quality improvement at an industry level is driven by removing the ability to hide technical inefficiency behind lower quality of care. Whether net benefit maximisation can be reached, within current cost per admission for any given clinical activity, as described in 6.4, and illustrated in 6.5, theoretically depends on the degree of technical, versus allocative, inefficiency, under current case-mix funding. In practice, it also depends on the ability to create accountability with peer identification, relative performance measurement and valuing effects of care in funding, conditional on effects of care, relative to best practice. Whether the scheduled value reaches a net benefit maximising level, within current case-mix funding levels, or not, moving from a zero to positive value for quality begins to create accountability for effects of care. Quality of care should improve in response to the incentives and accountability this creates. A hospital funding system with higher quality and the same cost per admission has the potential to be cost saving in treating patient populations within the health system across time, except potentially where patients die from lower quality of care. Any such cost savings, post-separation, could be seen as part of the buffer payment to increase scheduled value, towards net-benefit maximising levels.

In evaluating relative hospital performance, application of the correspondence theorem satisfies Ockham's Razor by being both simpler, and allowing greater explanatory power, than alternative methods specifying disutility events as outputs to represent quality of care, such as the hyperbolic approach of Färe, Grosskopf, Lovell and Parsuka (1989). While input specifications of disutility events have been previously employed to estimate technical efficiency (Pittman, 1981; Morey, 1992 and Rheinhardt, Lovell and Thijjsen, 1999), the correspondence theorem establishes that these frontiers also allow appropriate measurement of economic efficiency, where net benefit maximisation per service is an appropriate objective. The method provides policy makers with ability to estimate

economic and allocative efficiency, in addition to technical efficiency, and shadow price in the absence of output prices.

The main original contributions of this thesis include:

1. Identifying and illustrating the applicability of a problematic aggregation paradox (Fox, 1999) in current aggregate economic efficiency measures, such as cost per case-mix adjusted separation. In any bilateral comparison, a public hospital can be more economically efficient in each activity but be measured as less economically efficient at an aggregate level, due to exogenous patient mix. Despite perceptions to the contrary, case-mix adjusted cost per separation, while weighting admissions as outputs by relative industry average costs, does not allow for differences in cost share by activity in measuring aggregate efficiency. Aggregating economic efficiency at a clinical activity (DRG) level, using industry cost shares, was identified as providing a non-paradoxical aggregation method, while retaining relative industry importance of activities.
2. Deriving the net benefit correspondence theorem and illustrating its application as a method for including disutility events as effects of quality of services in performance measurement, consistent with net benefit maximisation.
3. Identifying and illustrating a 2-stage sequential funding mechanism that can maintain current payment levels per admission, while allowing a managed transition, from economic incentives for cost minimising care under case-mix funding, towards accountability and incentives consistent with evidence-based medicine in practice.
4. In health technology assessment, illustrating distinct advantages of incremental consideration of frontiers in the cost disutility plane, over frontiers in the incremental cost effectiveness plane. In addition to identification of dominance, and regions of willingness to pay for best practice, frontiers in the cost-disutility plane were illustrated in chapter 8 to allow estimation of degree of dominance with radial contraction, and explicit representation of differences in net health benefit and monetary benefit.

5. Identifying and illustrating a method to increase the precision of bootstrapped incremental cost-effectiveness ratios, estimated with patient level data from randomized control trials. Structural uncertainty was identified as inadvertently imposed in randomly matching re-sampled treatment and control patient populations with the current method (Briggs, O'Brien and Blackhouse, 2002) of bootstrapping the distribution of incremental cost effectiveness ratios. This structural uncertainty was illustrated, in appendix 8.1, as able to be minimised with stratification and ordered matching on prognostic risks of cost and effect, following Eckermann and Kirby (2003).

In considering further research, the correspondence theorem allows use of any appropriate method that can minimise costs and disutility events valued at WTP per service. To allow for potential effects of sample variation on frontiers, future research is suggested on the use of stochastic frontier analysis and use of Bayesian shrinkage estimation methods (Coory and Gibberd, 1998; Howley and Gibberd, 2003), prior to estimating frontiers with DEA in addition to adjustment for patient risk factors (Simpson et al., 2003). In addressing the primary question of appropriate economic incentives for quality of care, while hospital provider uncertainty at a clinical activity (DRG) level may be significant, the importance of uncertainty diminishes in aggregation across clinical activities and across time. Provided correspondence theorem assumptions have been satisfied, with differences in risk factors in patient populations adjusted for and effects measured within- and beyond-care measured at a clinical activity level, hospitals need only fear performance measurement and funding under the correspondence theorem if they have hidden technical efficiency behind lower quality of care.

The generalised nature of the correspondence theorem also suggests potential application in other service industries, where maximising net benefit per service is an appropriate objective, and quality of service can be represented by reduction in rates of disutility events. Maximising net benefit per service will in general be an appropriate objective for industries with quality differentiated services, where effects of services on disutility event rates are incremental and transaction conditions do not permit preferences for quality of

services to be inferred from consumer behaviour, or prices. Potential applications in public services (where prices do not exist and transaction conditions would not allow them to represent quality if they did) include other health care services, corrective services with disutility events such as recidivism, and employment placement services with disutility events of rates of placement failure and retention. In comparison of relative performance in these industries, as in hospitals, adjusting for risk factors of clients and effects beyond service are important to avoid incentives for cream-skimming and cost-shifting beyond service. The ability to frame disutility rates as average disutility, limitation of functioning or abilities, or where standards of service are not met, suggests the scope for application may be wider still.

In conclusion, in addressing the question of how to include quality of care represented by disutility events, this thesis has identified and illustrated a method for including disutility bearing effects as quality of care indicators in performance measurement, and funding mechanisms, consistent with maximising net benefit. In public hospitals, maximising net benefit per admission is an appropriate economic efficiency measure that, unlike cost effectiveness or cost minimisation, allows for the derived nature of demand for health care services, and the incremental and non-traded nature of health effects. In trading off the value and costs of quality of care, an underlying objective of net benefit maximisation, rather than cost minimisation, in funding and performance measurement, also creates a more appropriate internal negotiation process and dialogue between hospital administrators and health care providers (Harris, 1977) in allocating resources. Clinicians, by having the value of their quality of care recognized, become economically accountable for this quality of care. Administrators no longer act as accountants, minimising cost per admission relative to average industry levels, but rather are required to consider trade-offs between the value and cost of quality of care in maximising net benefit, relative to best practice.

In addressing questions of adjusting for risk factors and effects beyond separation, the correspondence theorem underlying the proposed method allows a synthesis of evidence

of pre-care risks and post-care effects with its linear form. In satisfying correspondence assumptions this method also explicitly identifies the need to adjust for these differences.

In relation to questions of appropriate level of analysis, to satisfy correspondence theorem conditions, a clinical activity level allows flexible and comprehensive identification of effects of care using decision analytic methods, and within-DRG patient risk factors to be adjusted for without confounding at aggregate levels. Incentives for cost-shifting and cream-skimming are then avoided in performance measurement and funding, in addition to providing incentives for quality of care consistent with maximising net benefit. A clinical activity level has also been shown to allow an aggregation paradox (Fox, 1999) to be overcome and processes of clinical audit and peer review to be employed in verifying and monitoring effects.

The correspondence theorem offers a systematic framework to create incentives for evidence-based medicine in a continuum from health technology assessment to performance measurement and funding. Satisfying explicit correspondence theorem assumptions of a common comparator and coverage of effects of care, unlike current performance measurement and funding frameworks provides natural robustness conditions in application. To satisfy these assumptions in practice a four stage process has been suggested in applying the correspondence theorem: (1) identification of effects of care at a clinical activity (DRG) level using decision analytic methods; (2) measurement of effects of care, including data linkage to effects beyond separation (preventing cost shifting incentives); (3) reframing effects as disutility event rates; (4) adjustment for differences in patient risk factors (to prevent cream skimming incentives).

Provided correspondence conditions are satisfied, limitations of this framework can be seen as those of maximising net benefit per admission as an appropriate objective function. In public hospitals the incremental and non-traded nature of health effects suggests advantages as in health technology assessment of maximising net benefit over minimising cost per unit effect.

Appendix 4.1: Data Envelopment Analysis (DEA): origins & formulations

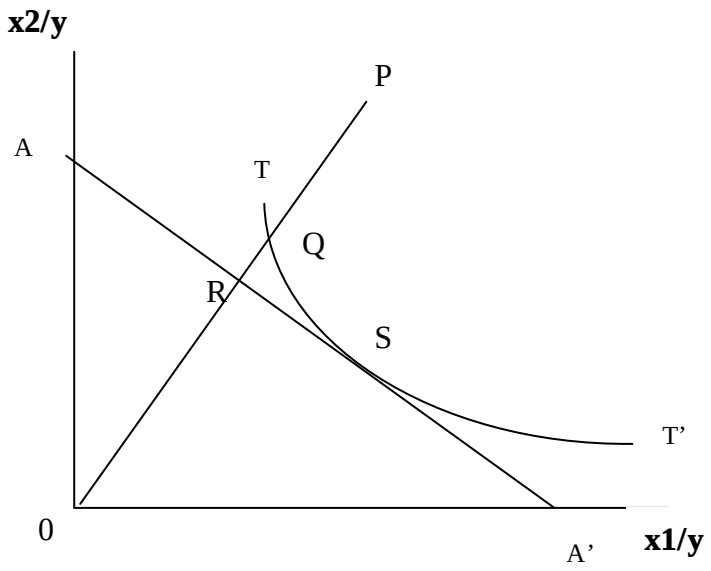
Farrell (1957) proposed that the efficiency of a firm consists of two components:

1. technical efficiency (TE), which reflects the ability of a firm to obtain maximum output from a given set of inputs (output orientation) or minimise inputs for a given level of outputs (input orientation) and;
2. allocative efficiency (AE), which reflects the ability of a firm to use inputs in the correct proportions given their relative prices and production technology.

From an input orientation these two measures are then combined to give cost efficiency (often called total or economic efficiency, but economic efficiency depends on the objective function, which may not be cost minimisation).

Farrell demonstrated that from an input orientation, under constant returns to scale, technical efficiency for an inefficient firm, at P in figure A4.1.1, can be measured relative to the unit isoquant (TT') as OQ/OP. Technical inefficiency is measured as PQ/OP, the proportion by which all inputs (x1, x2) can be reduced without changing outputs (y) in radially contracting on the unit isoquant.

Figure A4.1.1: Farrell technical economic and allocative efficiency on the unit isoquant



If the input price ratio is known (represented by the slope of the isocost line AA' in figure A.4.1.1), and cost minimisation is an appropriate objective, then from an input perspective, allocative efficiency of production can be measured for a firm at P as OR/OQ. Allocative inefficiency, RQ/OQ in figure A.4.1.1, represents the proportion by which production costs at technically efficient point Q could be reduced if inputs were used in the same proportions as at technical and allocative efficient point S.

Total cost efficiency for a firm at P can be estimated by the ratio OR/OP, where RP/OP represents the proportional cost reduction per unit possible.

Technical, allocative and cost efficiency measures assume the production function is known. In practice this is not the case and the efficient isoquant must be estimated by sample data. Farrell (1957) suggested either:

(a) a non-parametric piece-wise linear convex isoquant, constructed such that no observed point lies to the left or below it, or;

(b) a parametric function such as the Cobb-Douglas form fitted to the data, again so that no observation lies to the left or below it.

The first option is what was called Data Envelopment Analysis (DEA) by Charnes, Cooper and Rhodes (1978). Charnes, Cooper and Rhodes (1978) reformulated Farells' (1957) concept of a radial contraction onto a piecewise linear convex hull into a mathematical programming problem allowing the calculation of efficiency scores for individual decision making units.

DEA involves the use of linear programming methods to construct a non- parametric piece wise surface (or frontier) over the data representing best observed practice. Efficiency measures are then calculated relative to this surface. Following Coelli, Rao and Batesse (1998) we first describe the method under constant returns to scale.

The constant returns to scale (CRS) ratio form of the DEA problem formulation (A.4.1.1) tries to obtain for each firm (i=1 to n) with m inputs and k outputs a ratio of outputs (vector y) over inputs (vector x):

$$\max_{u,v}(u'y_i / v'x_i)$$
$$st$$
$$u'y_j / v'x_j \le 1, \qquad \text{(A4.1.1)}$$
$$j = 1,2,...n$$
$$u, v \ge 0;$$

where u is a m×1 vector of output weights and v is a k×1 vector of input weights.

Formulation (A4.1.1) involves finding values for u and v such that the efficiency measure for the ith firm is maximised, subject to the constraints that all efficiency measures must be equal to or less than 1. There are n constraints and an infinite number of solutions.

Adding the extra constraint, $v'x_i = 1$

the problem becomes:

$$\max_{u,v}(u'y_i)$$
$$st$$
$$v'x_i = 1, \qquad \text{(A4.1.2)}$$
$$u'y_j - v'x_j \le 0$$
$$j = 1,2,...n$$
$$u, v \ge 0,$$

This formulation (A4.1.2) is known as the multiplier form of the DEA linear programming problem, and has n+1 constraints. Using duality in linear programming, one can derive an equivalent envelopment form of this problem:

$$\min_{q,l} q$$
$$st$$
$$-y_i + Yl \ge 0 \qquad \text{(A4.1.3)}$$
$$qx_i - Xl \ge 0$$
$$l \ge 0$$

where θ is a scalar and λ is a n×1 vector of constraints.

The envelopment form involves fewer constraints than the multiplier form (k+m<n+1) and hence is the preferred form. The linear programming problem needs to be solved n times, once for each DMU (i=1 to n). The value of θ obtained in each of these n

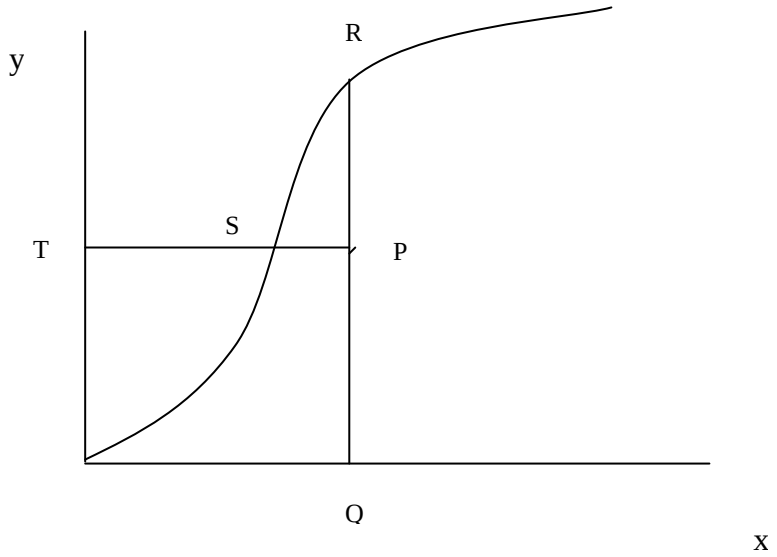
programming programmes is the efficiency score for the ith firm. It will lie between 0 and 1, with 1 indicating it lies on the frontier and hence a technically efficient firm according to the Farrell (1957) definition.

The intuitive interpretation of this envelopment form of the DEA problem from an input orientation is to take the ith decision making unit (hospital in our case) and seek to radially contract the input vector x_i as much as possible, while still remaining within the feasible input set. The inner boundary of this set is a piece-wise linear isoquant determined by the observed data points (all firms in the sample). The radial contraction of the input vector x_i produces a projected point ($X\lambda$, $Y\lambda$) on the surface of this technology. The projected point is a linear combination of observed best practice, with the constraints ensuring the projection cannot lie outside the feasible set.

Disaggregating technical efficiency under constant returns to scale: scale efficiency and technical efficiency under variable returns to scale

The formulation of DEA in each of the formulations A.4.1.1 to A.4.1.3 assumes constant returns to scale in calculating technical efficiency. Where production is characterised by constant returns to scale, such as that implicit in using a unit isoquant, input measures of technical efficiency are equivalent to technical efficiency from an output orientation. That is, with constant returns to scale, a radial increase in all outputs given inputs is the same as radial contractions in all inputs given outputs. However, if technology has variable rather than constant returns to scale, then input and output orientated efficiency scores will in general not be the same. This is illustrated in diagram A.4.1.2, in the single input (x), single output (y) case with increasing then decreasing returns to scale.

Figure A4.1.2: Orientation of technical efficiency where technology is not constant returns to scale



Under technology without constant returns to scale, such as the increasing an decreasing returns to scale in the production function OA in figure A4.1.2, technical efficiency will generally be different if measured with an input or output orientation. For an inefficient firm at P, technical efficiency under the Farrell definition with an input orientation would be measured as TS/TP and from an output orientation QP/QR.

Orientation of efficiency measurement is, therefore, important in measuring technical efficiency if production technology is unknown or known to not support constant returns to scale. To determine scale efficiency a variable returns to scale (VRS), DEA model can be used, provided the assumption of variable returns to scale is appropriate. Technical efficiency derived under constant returns to scale can be decomposed into technical efficiency under variable returns to scale (VRS), and scale efficiency estimated as a residual. Whether scale inefficiency is due to increasing or decreasing returns to scale can also be estimated from comparing technical efficiency under VRS with technical efficiency under non-increasing returns to scale (NIRS). However, it should be noted that the implicit assumption of DEA that observed best practice represents technology becomes more tenuous where comparators are restricted to those of similar size implicit in the VRS formulation. In estimating technical efficiency

under variable returns to scale and consequently scale efficiency, the assumption of best practice representing technology is effectively required across all scales of production.

DEA Variable Returns to Scale (VRS) and Scale Efficiencies (SE)

When not all firms are operating at the optimal scale, the CRS specification of DEA, results in measures of technical efficiency, which are confounded by scale efficiencies. Banker, Charnes and Cooper (1984) suggested an extension of the DEA model under CRS to allow for VRS. The use of the VRS specification permits the calculation of technical efficiency without scale efficiency effects.

Adding the convexity constraint N1'λ=1 constraint to the CRS formulation (A4.1.3) yields a VRS formulation (A4.1.4):

$$
\begin{aligned}
&\min_{q,l} q \\
&st \\
&-y_i + Yl \geq 0, \\
&qx_i - Xl \geq 0, \\
&N1'l = 1, \\
&l \geq 0.
\end{aligned}
\qquad \text{(A4.1.4)}
$$

The convexity constraint ensures that a DMU is only benchmarked against firms of a similar size. That is, the projected point on the DEA frontier will be a convex combination of observed firms. This approach forms a convex hull of intersecting planes that envelope the data points more closely than CRS, providing technical efficiency scores greater than or equal to that under the CRS DEA form. This compares with technical efficiency calculated with DEA under CRS where a firm may be compared against larger or smaller firms, if the λ weights sum to a value greater than or less than one respectively.

Given that one believes that the technology is VRS, technical efficiency scores under CRS can be decomposed into pure technical efficiency (VRS) and scale efficiency (SE): $TE_{crs} = TE_{vrs} \times SE$.

One shortcoming of this measure of scale efficiency is that the value does not indicate whether the firm is operating in an area of increasing or decreasing returns to scale.

This can be determined by running a non-increasing returns to scale (NIRS) DEA formulation:

$$\begin{aligned}
&\min_{q,l} q \\
&st \\
&-y_i + Yl \geq 0, \\
&qx_i - Xl \geq 0, \\
&N1'l \leq 1, \\
&l \geq 0.
\end{aligned} \qquad \text{(A4.1.5)}$$

The constraint N1'$\lambda \leq 1$ ensures the ith firm will not be benchmarked against larger firms, but may be compared with smaller firms. Therefore if $TE_{nirs} \neq TE_{vrs}$, the firm is in an area of increasing returns to scale, otherwise there are decreasing returns to scale with scale inefficiency present ($TE_{nirs} = TE_{vrs}$).

Allocative efficiency

If price information is available and a behavioural objective such as cost minimisation is appropriate, then it is possible to measure allocative, as well as technical, efficiency. To achieve this two sets of linear programming are required: one to measure technical efficiency and the other to measure cost (economic) efficiency shown in the linear program DEA formulation A4.1.7.

$$\begin{aligned}
&\min_{l,xi} ..w_i' x_i^*, \\
&st \\
&-y_i + Yl \geq 0, \\
&x_i^* - Xl \geq 0, \\
&N1'l = 1, \\
&l \geq 0
\end{aligned} \qquad \text{(A4.1.7)}$$

where w_i is a vector of input prices for the ith firm and x_i * (which is calculated by the LP) is the cost minimising vector of input quantities for the ith firm, given the input prices w_i and the output levels y_i. The total or cost efficiency (CE) of the ith firm is then calculated as:

$CE_i = w_i x_i^* / w_i x_i$.

That is, as the ratio of minimum cost to observed cost. Allocative efficiency (AE) can then be estimated residually as:

$$AE_i = CE_i / TE_i .$$

This procedure implicitly includes any slacks into the allocative efficiency measure. This is justified on the grounds that slacks represent inappropriate output mixes (Ferrier and Lovell (1990:235)). However, in the absence of economic and allocative efficiency measures slacks should be reported with technical efficiency.

Congestion efficiency

Where there is congestion in use of an input an isoquant may bend backwards and consequently have a positive slope. This corresponds to the declining part of the total product curve, where an input can have negative marginal product. While this should not arise with free choice of input combinations and appropriate incentives for decision making units, constraints on inputs, regulation or perverse incentives can cause this.

To allow for congestion in estimating technical efficiency DEA an assumption of strong disposability in inputs (costless disposal of inputs) implicit in DEA formulations A.4.1.1 to A.4.1.5 can be relaxed to allow weak disposability of inputs. Weak disposability of inputs is achieved by changing input constraint inequalities to equalities and adding a parameter in this restriction, as described in Fare, Grosskopf and Lovell (1985):

$$\begin{aligned}
&\min_{q,l,d} q, \\
&st \\
&-y_i + Yl \geq 0, \\
&dqx_i - Xl = 0, \\
&N1'l = 1, \\
&l \geq 0, \\
&0 < d \leq 1.
\end{aligned} \qquad \text{(A4.1.6)}$$

This formulation for technical efficiency under variable returns to scale, with weak disposability of inputs, allows a backward bending frontier, reflecting the potential for a negative marginal product of inputs beyond a certain point. Technical efficiency with strong disposability of outputs can, therefore, be decomposed into technical efficiency under weak disposability of inputs and congestion efficiency as a residual. However,

unless there is good reason to believe that congestion exists for inputs and is due to factors outside the control of DMUs, any estimated backward bending part of the frontier implicit in technical efficiency under weak disposability reflect poor choice of input factor proportions or technical inefficiency and cannot be regarded as best practice.

Appendix 4.2 Measuring technical efficiency with cost data

The National Hospital Cost Data Collection (CDHCS) provides cost data, but not quantity data, by factor input at a DRG level. In the absence of physical input or relative input price data, if only costs data by input is available, input orientated technical efficiency can still be estimated by DEA using the radial property of distance functions (Farrell 1957, Shephard 1970) under the constraint of an additional assumption. This appendix outlines the adaptation of DEA methods and assumption required to accommodate input costs data by factor rather than quantities in estimating technical efficiency.

As noted in appendix 4.1, DEA has its origins in, and is largely based on, distance functions (Farrell, 1957; Shephard, 1970) and the associated notions of technical and allocative efficiency. The conventional formulation of DEA requires quantities of inputs and outputs. However, noting that one of the radial properties of distance functions is unit invariance, costs of inputs can be used to replace input quantities under the assumption that all compared firms (hospitals) face the same prices for factor inputs.

From an input orientation, Farrell (1957) technical inefficiency is estimated as the amount all inputs could be proportionally reduced without a reduction in output. Technical efficiency is measured along rays form the origin and, as such, is described as a radial measure, which is unit invariant. Changing physical units for each axis by multiplying by a scaler, such as a common price, will not alter Farrell (1957) technical efficiency measured using radial contraction with distance functions (Shephard, 1970). This is the case, as both the physical inputs and prices are proportionally reduced in radial contraction.

Therefore, assuming all firms have the same prices for inputs, radial contraction in proportionately reducing all costs to a cost surface (comprising the best practice firms who are assumed to face the same prices) is the same as radial contraction in proportionately reducing all inputs to a physical input best practice surface.

If the assumption of hospitals facing the same input prices is violated, then measured technical efficiency will include price differences between hospitals for input factors. Where prices systematically differ, and that systematic difference is quantifiable between hospitals (as in differences between state awards for nurses), relative factor price differences could, however, be used to deflate between hospitals facing different prices in calculating technical efficiency. Price differences due to scale purchasing effects could also be minimised with technical efficiency calculated under variable returns to scale (VRS). Under VRS hospitals would be compared to those of similar size in estimating technical efficiency.

However, it should be noted that various biases can influence efficiency analysis based on costs. As Eckermann (1994) noted in critically appraising the use of cost per case-mix adjusted separation, estimating productivity using a measure of cost per output requires the assumption that factor prices such as wages reflect marginal product. Violations of this assumption can be particularly serious where factor markets in hospitals are not competitive. In Australia this may be particularly the case in the labour market for public hospitals.

Labour markets can be highly regulated with wages set by arbitration rather than a market. In relative performance measurement at a point in time, this can be problematic where unit prices differ across regions, and hospitals, due to relative bargaining power, or differences such as costs of living. These sources of biases in estimating productivity and technical efficiency are, however, also present in current comparison of cost per case-mix adjusted separation and in general can only be overcome if factor inputs rather than cost of inputs were available for use.

Economic efficiency and residual allocative efficiency with cost data alone

From an input perspective, economic efficiency with cost data by input for each firm is trivial as costs can be added, effectively ascribing relative prices of 1 to costs for each input included in technical efficiency. As with technical efficiency, this assumes that all hospitals face the same factor prices. The vector of input factor prices for the ith firm in the economic efficiency linear programming DEA formulation for economic efficiency (A.4.1.6 in appendix 4.1), is assumed to be a common vector across all firms. Minimising the product of factor inputs and prices faced by the ith firm, given output

and these factor prices, can be replaced by minimising cost given output of the ith firm, under the assumption of common relative prices.

Using distance functions under the assumption of firms (hospitals) facing the same prices for factor inputs, both technical and economic efficiency can therefore be estimated using output and cost data by factor input. Hence allocative efficiency can also be estimated as a residual of economic and technical efficiency.

An additional implicit assumption with use of cost by factor is that input cost factors contain the same input across hospitals in distinguishing between technical and allocative efficiency. In practice, this assumption of homogeneity of inputs within an input would also be required if physical inputs were used. However, care must be taken that the ability to aggregate costs does not mask factor differences. If costs for a given factor input contain multiple inputs, they need to be in the same proportions for all firms implying the same price index can be created as a scalar. To allow unbiased estimates of technical efficiency with cost data in practice, accounting practices need to be uniform in allocating costs to type of input in addition to hospitals facing common factor prices.

Appendix 5.1: Tobit regressions methods

To regress DEA generated technical efficiency (TE) scores against explanatory variables requires a method that can accommodate truncated data. Tobit regression first requires that DEA technical efficiency scores for each hospital are transformed by taking their reciprocal minus 1:

$$y_i = 1/TE_i - 1$$

Where TE_i is the DEA measure of DMU i's technical efficiency.

This transformation measures best practicing hospital as 0 and technically inefficient hospitals as an estimate of the relative radial distance between the efficiency frontier and the DMUs production point. Under constant returns to scale, this can be interpreted as the proportion by which all outputs could be increased without changing inputs.

Tobit regression on inefficiency as a dependant variable can then take the form:

$$y_i = b'x_i + m_i$$
$$if\ ..y_i > 0$$
$$0\ \text{otherwise}$$

where:

ß? a k× 1 vector of unknown parameters

x_i is a k× 1 vector of known constants (explanatory variables) for i.

y_i represents the transformed DEA scores for hospital i.

m_i are normally and independently distributed residuals, with mean 0 and common variance s^2.

The likelihood function L, which is maximised to solve for B and σ is then:

$$L = \prod_{yi=0}(1 - F_i)\prod_{yi>0} 1/(2\Pi)^{1/2} \times e^{-[1/(2s^2)](y_i - B'x_i)^2}$$

where

$$F_i = \int_{-\infty}^{B'xi/s} [1/(2\Pi)^{1/2} \times e^{-t^2/2}]dt.$$

As y_i is a measure of inefficiency, a positive estimate for a given ß coefficient suggests a factor is negatively affecting efficiency.

In the likelihood function the first product is over the observations for which DMUs are completely efficient and the second product where they are not. Fi is the distributional function of the standard normal, evaluated at B'xi/σ.

Testing the significance of the model can be undertaken with a Chi-squared test, based on a likelihood ratio, which tests the joint significance of the independent variables. The likelihood ratio is computed as −2log (LO/LI) where LI is the value of the likelihood model as fitted and LO is the maximum value if all coefficients except the intercept are 0.

Tobit regression used as a second stage regression on efficiency scores can be biased to the extent variables used in first stage technical efficiency measurement are correlated with second stage explanatory variables.

Appendix 7.1: Malmquist Index Methods

With access to suitable panel data of inputs and outputs frontier, Malmquist methods can be used to obtain estimates of total factor productivity growth without prices. A Malmquist Index of total factor productivity (TFP) uses distance functions, which do not require price data or the restrictive assumptions inherent in the Tornquvist/Fischer Index approach. Unlike the TFP indices Fischer/Tornquvist index approaches, the Malmquist TFP index of Caves, Christenson and Diewert (1982) can be decomposed into technical efficiency change (firms position with respect to the frontier) and technical change (shifting of the frontier). Färe et al.(1994) show how Malmquist component distance functions can be estimated using DEA.

The Malmquist Index does, however, make the strong assumption of constant returns to scale technology. Griffell-Tatje and Lovell (1995) demonstrate that, even with one input, one output and variable returns to scale, the Malmquist index may not correctly measure total factor productivity changes. Technology therefore needs to reflect a constant returns to scale technology to robustly interpret Malmquist Indexes.

From an input orientation, the input distance function defines the minimal proportional contraction of the input vector, given an output vector
The input distance function is defined on the input set L(y) as:

$$di(x,y) = \max\{\rho : (x/\rho) \varepsilon L(y)\}$$

The distance function di(x,y) is greater than or equal to one if the input vector, x, is an element of the feasible input set L(y). Further, the distance function will take a value of one if x is located on the inner boundary of the feasible input set, and will take a value less than one if x is located below the feasible input set.

Following Fare et al. (1994) the input orientated Malmquist TFP change index between period s (base period) and period t is given by:

$$m_i(x_s, y_s, x_t, y_t) = \{\frac{d_i^s(x_t, y_t)}{d_i^s(x_s, y_s)} \times \frac{d_i^t(x_t, y_t)}{d_i^t(x_s, y_s)}\}^{-1/2}$$

an equivalent way of writing this is:

$$m_i(x_s, y_s, x_t, y_t) = \frac{d_i^s(x_s, y_s)}{d_i^t(x_s, y_s)} \{ \frac{d_i^s(x_t, y_t)}{d_i^t(x_t, y_t)} \times \frac{d_i^s(x_s, y_s)}{d_i^t(x_s, y_s)} \}^{-1/2}$$

where:

$$\frac{d_i^s(x_s, y_s)}{d_i^t(x_s, y_s)}$$

is equivalent to technical efficiency change (that is the change in the input orientated measure of Farrell technical efficiency between periods s and t) and; .

$$\{ \frac{d_i^s(x_t, y_t)}{d_i^t(x_t, y_t)} \times \frac{d_i^s(x_s, y_s)}{d_i^t(x_s, y_s)} \}^{-1/2}$$

is a measure of technical change.

With appropriate panel data Malmquist indexes can be calculated using DEA to calculate each of the four distance functions for each of the firms as follows:

$$[d_i^t(x_t, y_t)]^{-1} = \min_{q,l} q,$$

$$-y_{it} + Y_t l \geq 0,$$
$$q x_{it} - X_t l \geq 0,$$
$$l \geq 0$$

$$[d_i^s(x_s, y_s)]^{-1} = \min_{q,l} q,$$

$$-y_{is} + Y_s l \geq 0,$$
$$q x_{is} - X_s l \geq 0,$$
$$l \geq 0$$

$$[d_i^t(x_s, y_s)]^{-1} = \min_{q,l} q,$$

$$-y_{is} + Y_t l \geq 0,$$
$$q x_{is} - X_t l \geq 0,$$
$$l \geq 0$$

$$[d_i^s(x_t, y_t)]^{-1} = \min_{q,l} q,$$

$$\begin{aligned} &-y_{it} + Y_s l \geq 0, \\ &qx_{it} - X_s l \geq 0, \\ &l \geq 0 \end{aligned}$$

As extra time periods are added, one must solve an extra three DEA linear programs. If there are T period of time then 3T-2 linear programs altogether need to be solved for each firm in the sample.

The above approach, if applied to time series data where input costs rather than physical quantities were measured, would additionally require that costs were deflated using implicit price deflators for years after the base year, prior to using them as input variables.

Appendix 8.1 Bootstrapping of the incremental cost effectiveness distribution allowing for baseline predictive factors

This appendix identifies and illustrates a method for increasing the precision of bootstrapping in estimating the uncertainty and confidence intervals of incremental cost effectiveness ratios (Eckermann and Kirby, 2003). Random matching of re-sampled treatment and control patients in forming incremental cost effectiveness ratio (ICER) bootstrap replicates, while not biasing the point estimate for the incremental cost effectiveness ratio, is shown to inadvertently introduce structural uncertainty. Differences in average prognostic risk of resource use and effects of care (e.g. myocardial infarction and cardiovascular mortality for the LIPID trial) between re-sampled treatment and control populations randomly matched are shown to be predictive of ICER. Undertaking stratified bootstrapping and ordered matching (based on average prognostic scores) between re-sampled populations by treatment arms, is shown to minimise differences between replicates in baseline risk, and increase precision of ICER distributions, relative to random matching. Stratification of re-sampling in bootstrapping and ordered matching of re-sampled populations in treatment arms based on prognostic risk improve precision of bootstrapping ICER confidence intervals, and the estimated bivariate distribution more generally. While improving precision these methods do not restrict other adjustment, such as use of the Kaplan Meier, and Kaplan Meier sample average method (Lin, Freur, Etzioni and Wax, 1997), to allow for censoring of patients effects and costs respectively.

Overview

Decision making based on incremental cost effectiveness ratios (ICERs) has been aided by recent developments in estimating ICER distributions from patient level data on costs and effects in randomised trials. Bootstrapping the ICER distribution based on re-sampling patients allows for covariance structure between costs and effects and the consequent ability to derive cost effectiveness acceptability curves (Briggs, O'Brien and Blackhouse, 2002). However, the random matching of treatment and control re-sampled populations to form ICER replicates, can introduce structural uncertainty, in not allowing for differences in re-sampled populations risks and hence reducing precision in estimating the ICER distribution. Differences in baseline risk between re-sampled populations by treatment arms do not reflect treatment effect. Such

differences are not permitted in decision analytic modelling of ICER distributions where Monte Carlo simulation applies common baseline risk across treatment alternatives (for example Eckermann, Martin, Stockler and Simes, 2003).

The LIPID study of secondary prevention of ischaemic heart disease with pravastatin (Glasziou, Eckermann, Mulray, Simes, Martin, Kirby et al., 2002) is used as a case study to illustrate the potential for structural uncertainty, and a method for minimising this.

The currently adopted method of bootstrapping the incremental cost effectiveness ratio distribution with implicit random matching of re-sampled treatment and control populations allows prognostic score differences between treatment and control populations in bootstrap replicates. In the LIPID study these differences are shown to be informative of ICER tail. Stratification and ordered matching by average prognostic score in bootstrapping the ICER distribution are demonstrated to minimise structural uncertainty inadvertently introduced in randomly allocating treatment and control populations to ICER bootstrap replicates. The increase in ICER precision from minimising this structural uncertainty is an empirical question. In the case of the LIPID study, precision of the 95% confidence interval for the cost per life saved is demonstrated to increase by 7.6%.

Background

Decision making based on cost effectiveness has been aided by recent developments in estimating uncertainty from randomised control trial data in bootstrapping (Briggs, O'Brien, and Blackhouse 2002). Bootstrapping of the ICER sampling distribution entails a four stage process:

1. Randomly sample with replacement (with sample size equal to the number of patients) individual patient cost and effect pairs from the control arm: calculate means for cost and effect.
2. Undertake step 1 for treatment arm/s.
3. Calculate a bootstrapped ICER replicate as the incremental average cost of treatment divided by incremental absolute effect (outcome) of treatment
4. Repeat steps 1-3 many times (at least 1000) to create the bootstrap estimate of the ICER sampling distribution

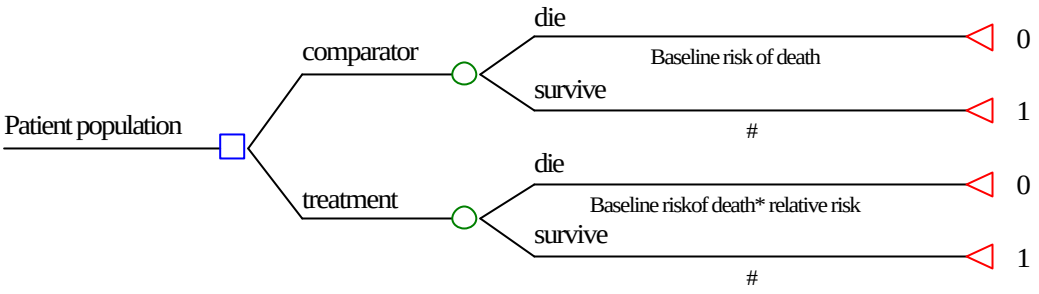
This method allows for covariance structure between costs and effects in estimating the ICER distribution and, as such, is preferable to a box method which considering costs and effects separately, in estimating 95% confidence intervals, and distributions more generally for the ICER. Additionally as the method identifies a bivariate distribution the method can also be used to estimate cost effectiveness acceptance curves (Briggs, 2002) and uncertainty of dominance (Briggs, Goeree, Blackhouse and O'Brien) or net benefit.

With complete follow up, the bootstrap method is only limited in modelling uncertainty of the ICER distribution for cost per life year saved in randomised control trials to the extent that cost and effect variables have been measured stochastically at a patient level for all patients.

While bootstrapping of the ICER does not require parametric assumptions with respect to bi-variate relationship between costs and effects, it is not, however, completely assumption free. In bootstrapping of the ICER distribution, estimation of mean costs and effect from patients occur separately in treatment and control arms under the assumption of independence from randomisation. However, even with independence and random allocation of re-sampled treatment arm population to bootstrapped ICER replicates, predictive baseline population differences in matched treatment and control populations will occur, whether or not, differences exist in the trial itself.

In implicitly undertaking random matching on re-sampled populations between treatment arms in bootstrapping the ICER, there is effectively an assumption that random differences in treatment populations by arm represent treatment effects. To the extent that these predictive differences in bootstrap replicates are informative of bootstrapped ICER tail distribution, uncertainty will be increased. Allowing differences in baseline risk by treatment arm, with randomly matched treatment populations in ICER, replicates is analogous to allowing baseline risk to differ by treatment arm in Monte Carlo simulations of ICER distributions. In Monte-Carlo simulation, based on decision analytic modelling, it is clear that such differences do not reflect uncertainty of treatment effect and should be removed by using common baseline risk across treatment strategies for any replicate (Figure A8.1.1).

Figure A8.1.1: Common baseline risk in decision analytic modelling



For example, in a study estimating uncertainty of incremental cost per quality adjusted life years saved of Tamoxifen used as a preventive in high risk women, Eckermann, Martin, Stockler and Simes (2003), applied common baseline risk of breast cancer in each in Monte-Carlo simulation of incremental cost per QALY saved.

Imbalance in predictive factors, between randomly allocated re-sampled populations by treatment arm in bootstrapping the ICER distribution, do not represent variation in treatment effect on cost effectiveness, while systematically widening the ICER confidence interval. Therefore, random allocation of replicates between treatment and control, while not affecting the expected ICER because of the random nature of re-sampling, will predictably introduce structural variation into estimation of the ICER distribution and 95% confidence intervals. The extent of this structural uncertainty will depend on how predictive differences in baseline risk are of the ICER ratio.

Methods

The LIPID trial of pravastatin use in secondary prevention of ischemic heart disease is used to illustrate methods for minimising structural uncertainty, inadvertently introduced with random matching of treatment and control populations, in bootstrapping incremental cost effectiveness ratios.

LIPID trial methods

The LIPID trial was a double-blind, randomised, placebo-controlled trial evaluating the long-term effects of pravastatin on all-cause mortality and coronary disease mortality in

patients who had had unstable angina or an acute myocardial infarction, and had a total cholesterol level of 4.0 to 7.0 mmol/L. The trial involved 9014 patients at 85 centres in Australia and New Zealand. Patients were given dietary advice conforming to the National Heart Foundation's recommendations and randomly assigned to 40 mg of pravastatin or placebo daily. The cost-effectiveness sub-study aimed to estimate the incremental cost per death averted within-study and the incremental cost per life year gained extrapolating within-study effects (Glasziou, Simes et al, 1997). Data on survival, medication and hospitalisation use were collected as part of the main LIPID study for all patients. Sample based sub-studies for diagnostic tests, ambulatory care and drug dose were collected, as well as a sub-study of nursing home use for stroke patients.

LIPID Analysis

All data were analysed on an intention-to-treat basis. Consistent with guidelines of the Pharmaceutical Benefits Advisory Committee (Commonwealth Department of Human Services and Health 1995), only direct (health system) costs were considered, to reflect the perspective of the health sector. Also consistent with PBAC guidelines, an annual 5% discount rate was applied to the incidence of costs and benefits, after baseline, to determine present value of cost per life year saved. Expected cost per life saved was calculated in 1997 Australian dollars, as the ratio of incremental costs (pravastatin less placebo) per person over the LIPID study period (average 6.0 years for survivors in both arms) to incremental reduction in all cause absolute mortality (placebo less pravastatin) over the study period. In LIPID all survival effects and over 97% of incremental cost effects were available at a patient level for all patients. The remaining 2.5% ($79/$3246) of incremental costs from sub-studies of ambulatory care ($45), diagnostic tests ($13) and nursing home care ($21) for stroke patients were modelled as fixed costs in estimating the bootstrapped distribution of the ICER.

Uncertainty of costs and effects in the LIPID study

Methods for estimating confidence intervals for incremental cost per life saved in the LIPID study were reported by Glassziou, Eckermann, Mulray, Simes, Martin, Kirby et al. (2002). Confidence intervals for incremental cost per death averted were calculated using the bootstrapping method, described by Briggs, O'Brien and Blackhouse (2002). 10,000 random re-samples with replacement were undertaken for 4502 patients in the

placebo and 4512 patients in the pravastatin study arm. These re-sampled populations were then randomly matched by treatment arm and their mean cost and survival rate used to form each ICER replicate, in the estimated sampling distribution.

Prognostic Index

For each LIPID patient, a prognostic score, representing risk of coronary heart disease (CHD) death or non-fatal myocardial infarction (MI) was calculated using the prognostic scoring index of Marschner et al. (2001). This index was derived from a multivariate risk factor model using Cox regression. Variables found to be significant in the model were: HDL cholesterol, age, gender, smoking status, qualifying event (Myocardial Infarction (MI) or Unstable Angina Pectoris (UAP)), diabetes, hypertension, prior revascularization and prior stroke.

For each individual a prognostic hazard ratio score (PHRS) was estimated as the additive sum across all factors of the log of hazard ratio attributable to each factor.

$$PHRS = \sum_{i=1}^{n} \ln(HR_i).$$

A higher PHRS indicates a higher risk of CHD death or non-fatal MI. For any treatment and control population combined to form an ICER replicate, a PHRS baseline treatment advantage is estimated as the average PHRS across placebo patients less the average PHRS for pravastatin patients. A positive PHRS baseline treatment advantage implies a higher average base risk in the placebo arm of CHD deaths and non-fatal MI. A negative PHRS treatment advantage implies patients with higher average risk in treatment arm of CHD deaths and non-fatal MI.

Hypothesis

PHRS treatment advantage is hypothesised to be informative for tails of the ICER distribution, with greater treatment advantage expected in lower tails than upper tails. With a null hypothesis that PHRS treatment advantage is not informative of tail distributions, average PHRS treatment advantage of replicates in the ICER upper and lower tails of the 95% ICER CI is tested for significant difference with a t-test adjusted for unequal variances.

Bootstrapping the ICER

To examine the effects of adjusting for prognostic factors in estimating the ICER distribution, four alternative methods of undertaking bootstrapping were compared:

1. bootstrapping without any adjustment for prognostic factors;
2. bootstrapping with stratification on prognostic factors;
3. bootstrapping with ordered matching by prognostic hazard ratio score;
4. bootstrapping combining stratification and ordered matching on prognostic hazard ratio score.

In the use of each of these methods 10,000 random samples with replacement were undertaken for 4502 patients in the placebo arm and 4512 patients in the pravastatin therapy arm.

1. Bootstrapping without adjustment for prognostic factors

Bootstrapping of the cost per life saved was initially undertaken with a four stage process as described by Briggs, O'Brien and Balckhouse (2002):

1. Randomly sample with replacement (with sample size equal to the number of patients) individual patient cost and effect pairs from the control arm: calculate means for cost and effect.
2. Undertake step 1 for treatment arm/s.
3. Calculate a bootstrapped ICER replicate as the incremental average cost of treatment divided by incremental absolute effect (outcome) of treatment
4. Repeat steps 1-3 many times (at least 1000) to create the bootstrap estimate of the ICER sampling distribution

This method implicitly randomly matches re-sampled (with replacement) pravastatin and placebo patient populations and their associated mean costs and effects. An incremental cost per life saved was estimated for each randomly matched placebo and control population (replicate) to estimate the ICER sampling distribution.

2. Bootstrapping with stratification

Stratified bootstrapping samples with replacement from patients within defined risk factor strata (for example the strata used in stratified randomisation), then combines re-samples from strata across the treatment population to estimate mean cost and effect by treatment arm. This adds a process in steps 1 and 2 of the four-stage bootstrapping of the ICER sampling distribution without stratification. In LIPID, age (31-64 years versus 65-75 years at study baseline) and gender were used to stratify bootstrapping, sampling with replacement within each age-gender strata in each treatment arm, then combining strata within each treatment arm.

3. Bootstrapping with prognostic score ordered matching

Prognostic score ordered bootstrapping of the ICER sampling distribution was undertaken in a six-stage process:

1. Randomly sample with replacement a sample of patients from the control study arm equal to the study population (4502 patients in the placebo arm in LIPID) with their associated costs and effects: calculate means cost and effect.
2. Undertake step 1 for treatment arm (a sample of 4512 patients in the pravastatin arm).
3. Repeat step 1 and 2 many times (10,000)
4. Order these 10,000 bootstrap replicates in each treatment arm by their average prognostic hazard ratio score.
5. Match the treatment and control bootstrap replicates based on the ordered in step 4.
6. Calculate the Bootstrap estimate of the ICER sampling distribution as the incremental cost of treatment, divided by incremental effect of treatment, from these 10,000 prognostic score ordered and matched ICER replicates.

For given treatment and control re-sampled populations, this method minimises variation in treatment score prognostic score advantage across ICER replicates.

4. Bootstrapping with stratification and prognostic score ordered matching

Stratified and prognostic score ordered bootstrapping combined patients re-sampled with replacement within strata bootstraps, as in age-gender stratified bootstrapping, then matched treatment arm mean costs and effect pairs, as for predictive score ordered bootstrapping of the ICER distribution.

Results: Illustrating the need for, and effect of, allowing for prognostic differences

Incremental cost effectiveness results from LIPID

The LIPID study of pravastatin use in the prevention of secondary ischemic heart disease (LIPID study group 1998) demonstrated a 3.0% absolute risk reduction in all cause mortality (95% CI 1.6-4.4%), at an average 6.0 years follow-up. Pravastatin therapy was associated with a significant relative risk reduction of 22% (95% CI 13 to 31%), and a baseline risk of 14.1% mortality in the placebo arm over the study follow up (figure A8.1.2).

Over the same average 6 year follow up period, incremental costs of pravastatin of $4913 per person were offset by reduction in other medication costs of $360 per person (95% CI $272 to $448) and hospitalisations of $1385 per person (95% CI $804 to $1966), resulting in total incremental costs of $3167 per person (95% CI $2559 to $3776). Incremental cost per life saved within the trial period was consequently estimated at $104,500 per life saved.

Using the box method of separately calculating confidence intervals for incremental effects and costs, bounds of between $58,000 per life saved and $236,000 per life saved were estimated from the extremities of ICER ratios in the box defined by 95% confidence intervals for costs and effects (figure A8.1.3).

Bootstrapping of the cost per life year saved with implicit random matching following Briggs, O'Brien and Blackhouse (2002), allowed an estimated 95 percent confidence interval from $67,000 to $204,000 per life saved, as shown in the cost effect plane in figure A8.1.4 and the cost effectiveness acceptance curve in figure A8.1.5.

Figure A8.1.2: LIPID study Kaplan Meier estimates of cumulative all cause mortality: pravastatin vs placebo.

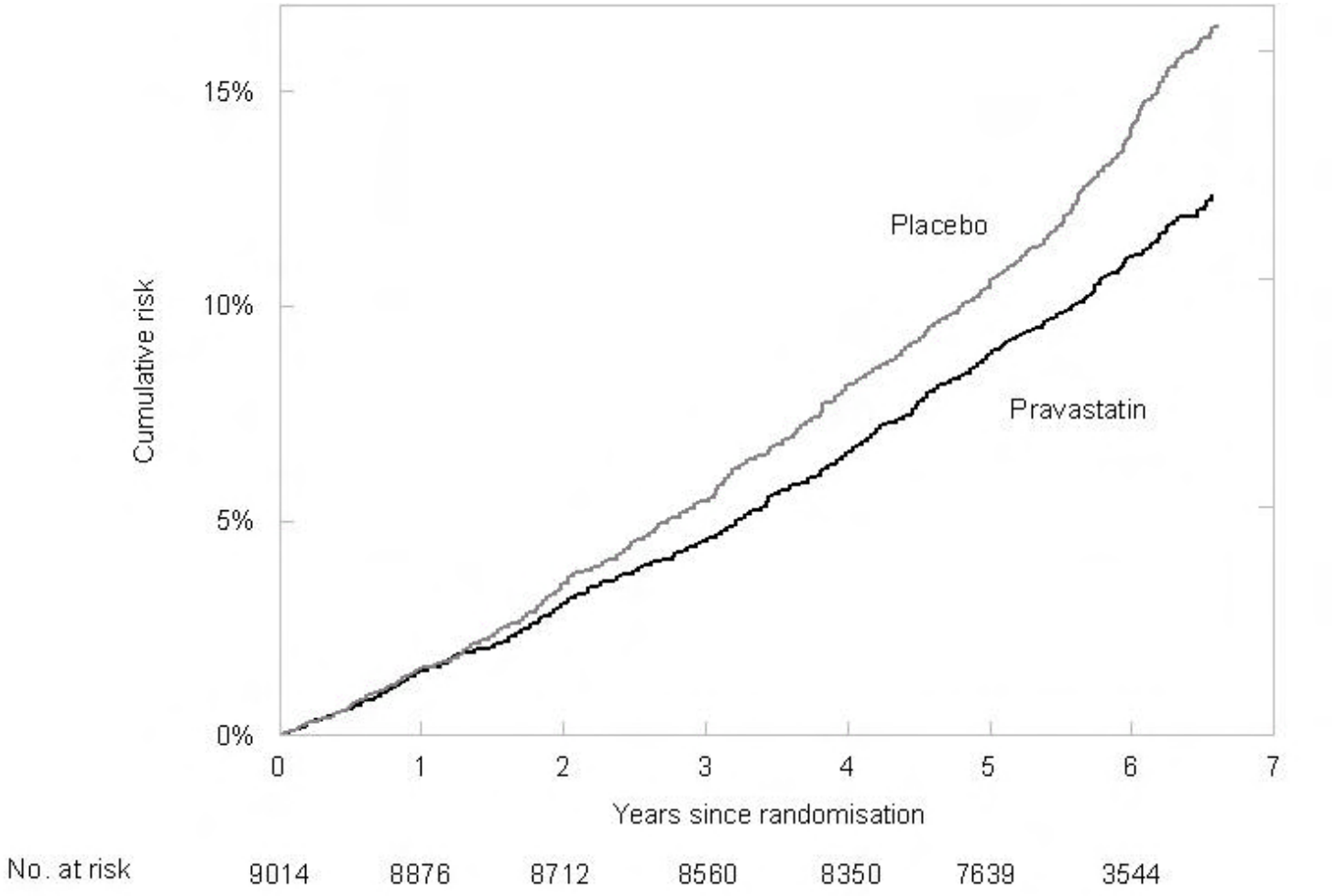


Figure A8.1.3: Box method for cost per life saved in LIPID (95% CI)

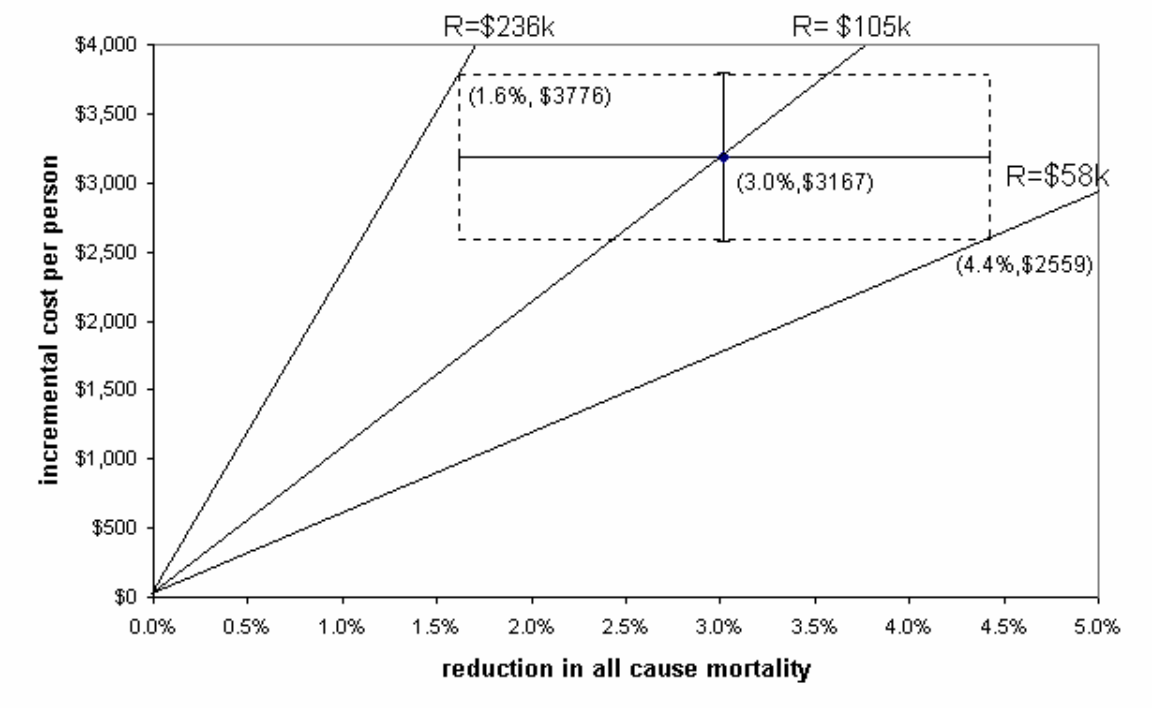


Figure A8.1.4: LIPID bootstrapped distribution (n=10,000) of cost per life saved with 95% confidence interval.

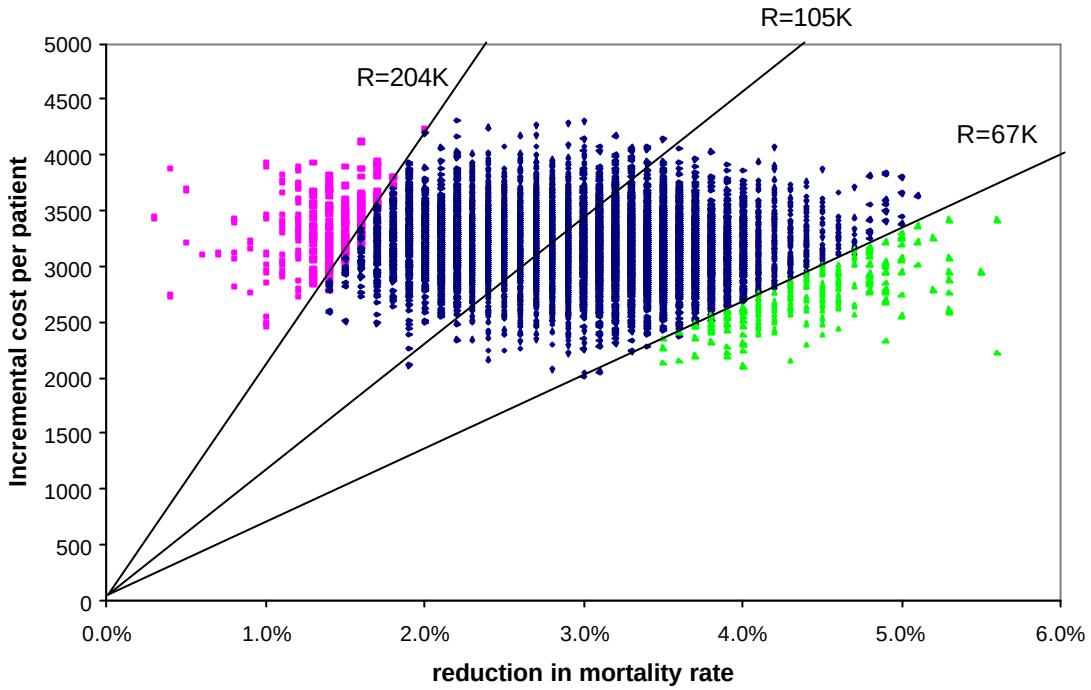
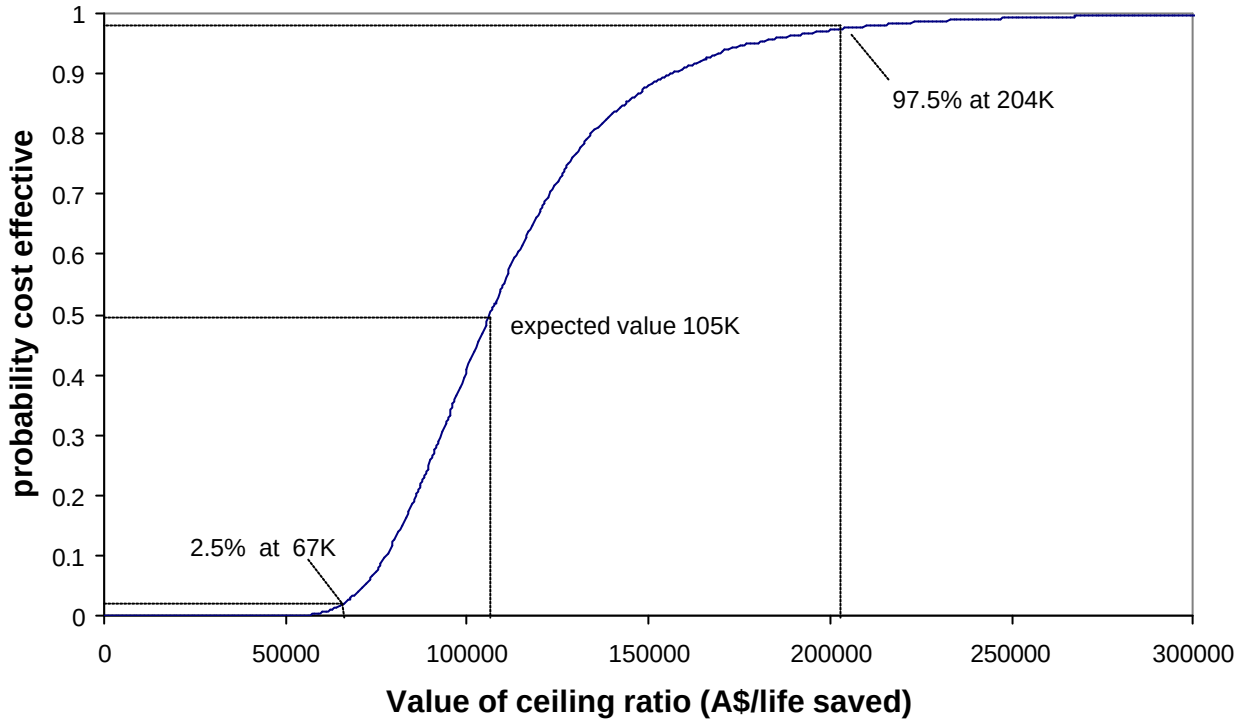


Figure A8.1.5: LIPID cost-effectiveness acceptance curve (probability cost effective conditional on threshold value of $ per life saved)



Balancing of predictive risk in bootstrap replicates

In LIPID, populations were well balanced between predictive factors by arm (LIPID study group, 1998), and there was no significant difference in average baseline prognostic factor score between pravastatin and placebo populations (p=.43). However, with random matching of re-sampled treatment and control populations, significant differences (at 5% level) emerged in the differences in prognostic hazard ratio scores between treatment and control population for 24% of replicates. In comparison, ordered matching between control and treatment re-sampled populations, based on average prognostic score, there were no significant differences between treatment arm predictive scores in any of 10,000 replicates. Ordered matching minimises differences between treatment and control for given re-sampled populations and hence the potential for introducing uncertainty related to prognostic differences. Stratification additionally allows the composition of re-sampled patient populations to be better balanced, as well as their average risk of modelled factors across patients.

While differences in prognostic score between re-sampled treatment and control populations in ICRE replicates are minimised by ordered matching, the further question is whether prognostic score differences are informative of ICER tail distributions. This is an empirical question, which is examined in testing whether there is a difference in predictive score in ICER tail distributions.

ICER relationship with prognostic score treatment advantage

The prognostic hazard ratio score (PHRS) treatment advantage from prognostic modelling in the LIPID study (Marschner et al. 2001), represents excess risk of CHD deaths or non-fatal myocardial in placebo relative to pravastatin re-sampled populations. In ICER bootstrap replicates under random matching, the PHRS treatment advantage is negatively associated with incremental cost (figure A8.1.6), positively associated with reduction in mortality (figure A8.1.7), as expected, and consequently negatively associated with incremental cost per life saved (figure A8.1.8).

Figure A8.1.6: Negative relationship between prognostic score treatment advantage and incremental costs per person over study (pravastatin less placebo) in bootstrap replicates with random matching

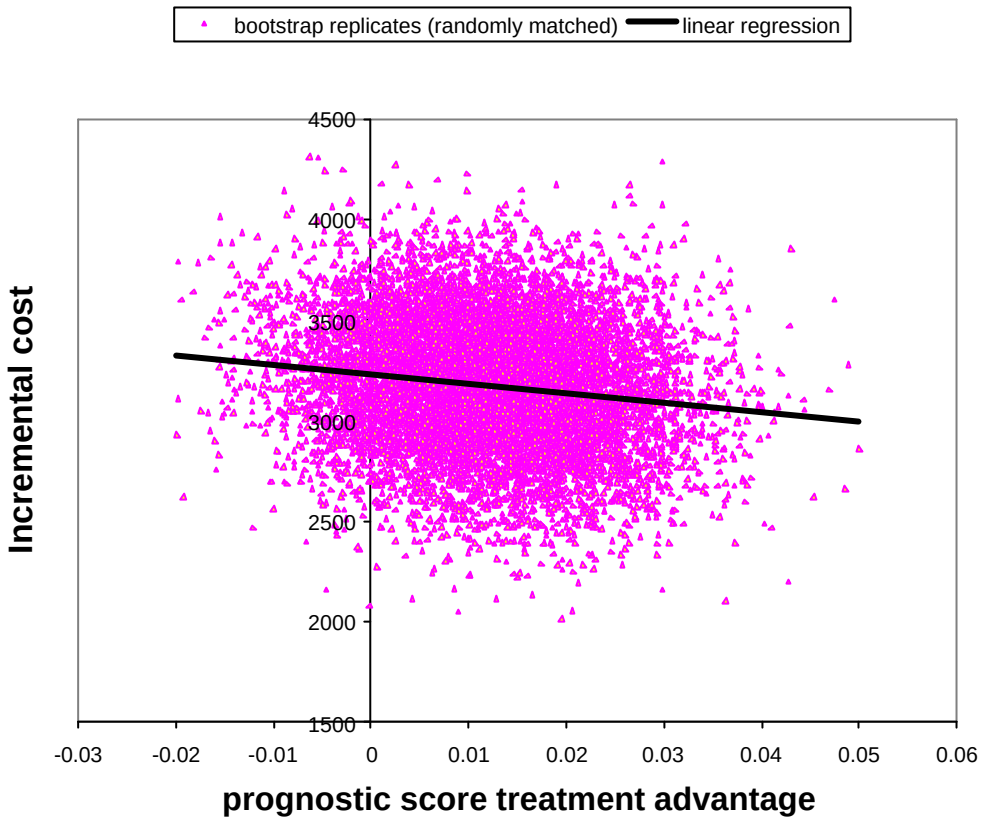


Figure A8.1.7: Positive relationship between prognostic score treatment advantage and absolute mortality reduction in bootstrap replicates with random matching

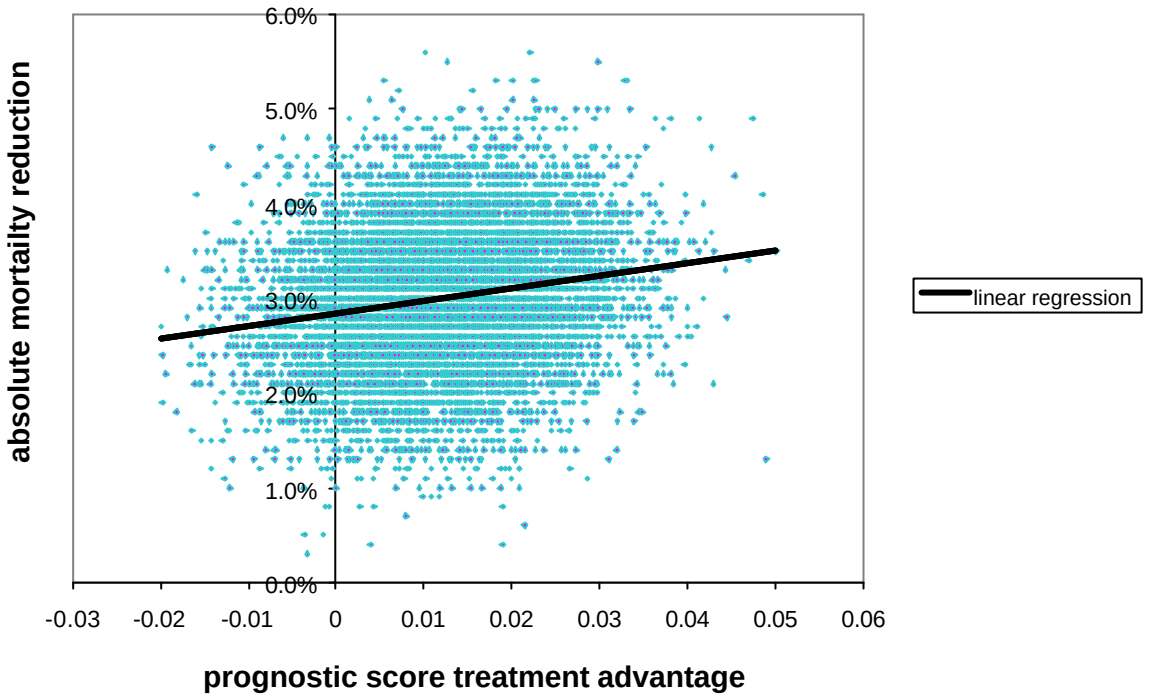
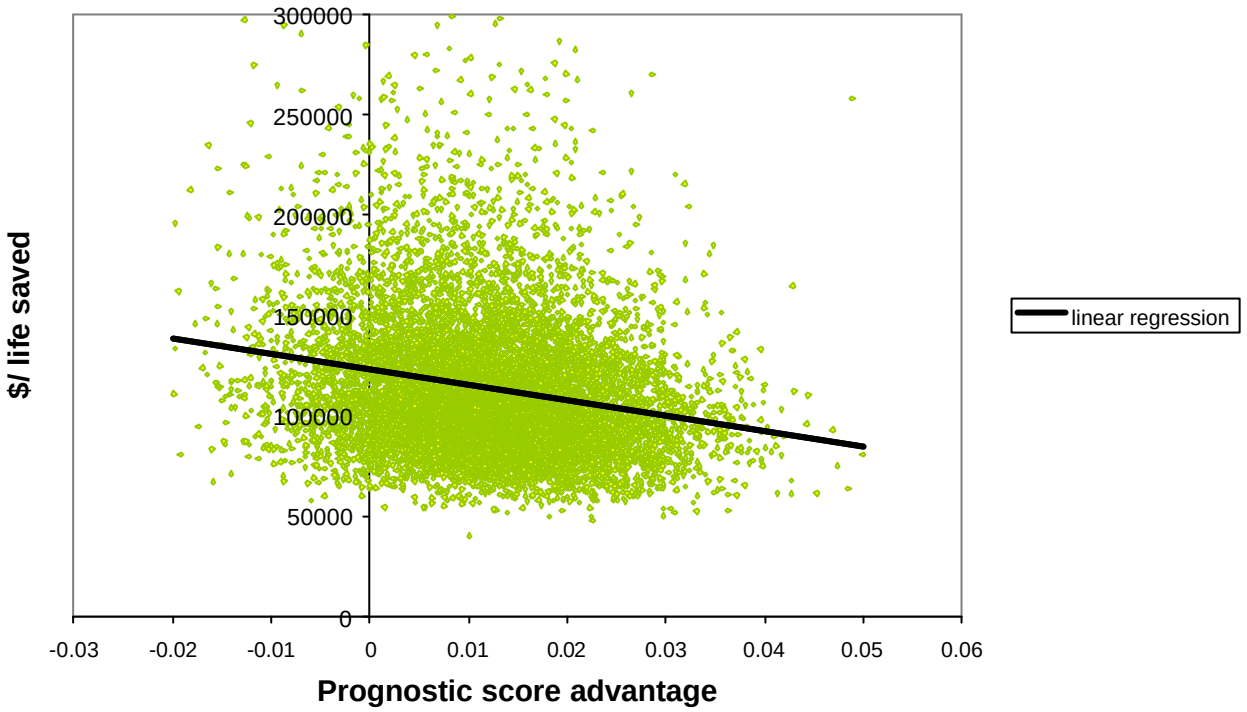


Figure A8.1.8: Negative relationship between prognostic score treatment advantage and incremental costs per person over study (pravastatin less placebo) in bootstrap replicates with random matching



The negative association of predictive treatment advantage with costs may be attributed to the presence of myocardial infarction in the prognostic index as a proxy for resource use, but, also, higher costs associated with deaths than survivors over the study period. The negative relationship between ICER and prognostic index treatment advantage reflects a negative relationship with cost, and positive relationship with treatment effects, reinforced by a general negative correlation between costs and survival.

Examining predictive score advantage in the tails of the bootstrapped ICER distribution, the average prognostic hazard score treatment advantage is significantly greater for cost per life saved ratios in the lower tail than the upper tail (mean 0.017 vs. 0.007, $p<.0001$). This supports the hypothesis that prognostic score differences at baseline are informative of tail distributions of the ICER, and with the expected sign. In comparison, with ordered matching, there is no difference in average treatment advantage, of 0.12 in upper and lower 2.5% tails. In minimising structural uncertainty, ordered matching on predictive factors, therefore, appears empirically, as well as theoretically, justified in LIPID. In table A8.1.1 the effects on estimated precision of ICER distributions (95% confidence intervals) from stratification, ordered matching and combined stratification and ordered matching are estimated.

Table A8.1.1: LIPID 95% confidence intervals for incremental cost per life saved by method of adjustment for prognostic risk in bootstrapping.

Predictive score matching	*95% CI limits*	*Bootstrap*	*Bootstrap with age & gender stratification*
No	Lower	$66,671	$66,794
	Upper	$204,071	$195,393
Yes	Lower	$67,614	$67,609
	Upper	$198,089	$192,227

Conditioning on predictive factors between pravastatin and treatment, by either ordered matching of bootstraps or stratification by age and sex, narrowed the 95% CI for cost per life saved. Lower 95% confidence interval bounds increased and upper bounds decreased in each case. Combining age-sex stratification and ordered matching

increased precision further. Empirically this suggests that, in the LIPID study, structural uncertainty from random matching between treatment arm populations, in bootstrapping the ICER, is present.

Stratification allows for the distribution of prognostic factors stratified, while ordered matching allows for variation within strata. Both appear important for LIPID, improving precision (as the ratio of upper and lower ICER 95% CI) by 4.4% with matching, 4.6% with stratifying and 7.6% in combination.

Discussion

Imbalances in predictive factors in randomly matched re-sampled populations in bootstrapping the ICER can lead to inadvertent introduction of structural uncertainty, if these imbalance are informative or predictive of ICER. The level and distribution of predictive imbalance will depend on population variability in predictive scores, in part a function of the size of study populations in each treatment arm. The degree of effect a given distribution of difference in predictive factors has on the ICER distribution will also depend on the strength of relationship between prognostic scores, costs and effects. The degree of improvement in precision in stratifying bootstrapping and undertaking ordered matching on prognostic risk scores will be greatest for randomised control trials with:

1. high variability in predictive score amongst participants in treatment arm/s, a function of sample size, as well as intrinsic patient heterogeneity and;
2. Prognostic index treatment advantage with strong negative relationship to incremental cost effectiveness.

The relationship between prognostic index treatment advantage (lower prognostic score in treatment arm) and ICER will be strongest where the prognostic index represents resource use as well as negative effects (morbidity, mortality). That is, where prognostic score treatment advantage is expected to be associated with a reduction in costs and reduction in negative health outcomes.

Combining ordered matching with other methods

Stratification and ordered matching do not impose any restrictions on what can be calculated in the bootstrapping of the ICER, while minimising variation in predictive

score differences between given re-sampled patient populations. Therefore, these methods can be used without constraint in combination with methods such as extrapolation, allowing for censoring in calculating incremental effects using Kaplan Meier and calculating costs using Kaplan Meier Sample Average methods (Lin, Freuer, Etzioni and Wax, 1997), or any other appropriate analysis.

If there is imbalance in predictive factors for effects in the actual trial populations, then Cox regression (Cox, 1972) could be considered appropriate as it can account for this baseline imbalance, while matching and stratification will not. In general, however, Cox regression requires severe truncation, does not allow freedom to additionally deal with censoring or extrapolation of effects in calculating the bootstrapped ICER distribution and does not account for predictive cost factors. Where imbalances exist in the study itself, direct adjustment of the linear net benefit statistic with individual patient data following Hoch, Briggs and Willan (2002) is likely to be more fruitful than approaches such as Cox Regression on effects alone.

Where imbalances do not exist across treatment arm in the study itself, stratification of bootstrapping on categorical risk factors and ordered matching on prognostic scoring models, based on categorical or continuous risk factors, is suggested. Depending on the nature and number of predictive cost and effect factors and available data, approaches such as stratification of bootstrapping on effect factors, and ordered matching by cost factors, such as time weighted cost histories, are suggested.

Feasibility of ordered matching approach

If a prognostic index for resource use and negative outcomes (mortality, morbidity) exists for identifiable patient characteristics, then the ordered matching by predictive factors in bootstrapping the ICER is simple. The method only requires a prognostic score for each individual and the ordering of treatment and control replicates by their average predictive score per patient. For large trials it has become routine for prognostic index specific to the trial to be calculated, and in general, if Cox regression is undertaken, a prognostic index can be constructed from placebo arm patients. For the purposes of ordered matching, a prognostic index need not, however, be trial specific. Pre-existing indices can be used, provided common covariates exist between

the study population and the index and the indexes endpoint represents risk of negative health events and resource use.

Conclusions

Random matching of re-sampled treatment and control populations in bootstrapping incremental cost-effectiveness ratios, while not biasing point estimates, has been demonstrated to introduce structural uncertainty in estimating the precision of the distribution. The extent of this is an empirical question, depending on the extent of variance in differences of prognostic risks of costs and effects and how predictive these differences are of incremental cost-effectiveness ratios.

To minimise structural uncertainty in undertaking bootstrapping of the ICER distribution, bootstrapping should at least mirror the clinical study in terms of stratification on predictive factors. Additionally, ordered matching on prognostic scores for resource use and negative effects (morality, morbidity) minimises imbalance in informative prognostic score differences between control treatment pairs. If baseline imbalances are also present, the approach of adjusting the linear net benefit statistic parametrically, with individual patient data, following Hoch, Briggs and Willan (2002), is likely to be more fruitful than Cox regression on effects alone.

In bootstrapping the ICER distribution with ordered matching on prognostic factors, predictive factors for costs, as well as clinical endpoints, should be considered in choosing endpoints and covariates for predictive indices use in ordered matching. The LIPID prognostic index (Marschner et al. 2001) provides an example of this in combining CHD mortality and non fatal MI endpoints. Stratification on categorical effects and ordered matching on cost histories, where available, as predictors of future costs, are suggested in further applications of the outlined approach.

References:

Aigner D., Lovell C. and Schmidt P. (1977). Formulation and Estimation of Stochastic Frontier Production Function Models. *Journal of Econometrics*;6:21-37.

Akerlof G. (1970). The Market for 'Lemons':Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics*;84(3):488-500.

Arrow, K.J. (1963). Uncertainty and the Welfare Economics of Medical Care. *American Economic Review*; 53: 941-973.

Arrow K.J. & Fischer A.C. (1974). Environmental preservation, uncertainty and irreversibility. *Quarterly Journal Of Economics* May 1974:312-320.

Australian Council on Healthcare Standards. (2001). Determining the Potential to Improve the Quality of Care in Australian Health Care Organisations: Trends in Quality of Care: Results from the ACHS Clinical Indicators data 1998-2000. ISBN 1 876987 02 2, 2001.

Australian Commonwealth Department of Health and Ageing Website. (2002). *2002 Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee including major submission involving economic analyses* http://www.health.gov.au/pbs/pubs/pharmpac/interim/index.htm. [11 December 2002].

Banker R.D., Charnes A., and Cooper R.C. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science*. 30: 1078-92.

Banker R.D., and Morey R.C. (1986a). The use of Categorical Variables in Data Envelopment Analysis. *Management Science* 32:1613-1627.

Banker R.D., and Morey R.C. (1986b). Efficiency Analysis for Exogenously Fixed Inputs and Outputs. *Operations Research*. 34: 513-521.

Banker R.D., Conrad R.F. and Strauss R.P. (1986). A comparative application of data envelopment analysis and translog methods: an illustrative study of hospital production. *Management Science* 32(1):30-44.

Birch S. and Gafni A. (1992). Cost effectiveness/utility analyses: do current decision rules lead us to where we want to be? *Journal of Health Economics* 11:279-296.

Birch S. Gafni A. (2002). On Being NICE in the UK: guidelines for technology appraisal for the NHS in England and Wales. *Health Economics* 11:185-191.

Blackhouse G., Briggs A. and O'Brien B. (2002) A Note on the Estimation of Confidence Intervals for Cost-Effectiveness When Costs and Effects are Censored. *Medical Decision Making* 22;2:173-177.

Boyce N., McNeil J., Graves D. and Dunt D. (1997). Quality and Outcome Indicators for Acute Healthcare Services. Commonwealth of Health and Family Services, Australia. Available online: http://www.health.gov.au/pubs/qualrprt/execsmry.pdf September 16 2003 date last accessed.

Boyle , K.J. and Bishop R.C. (1988) Welfare measurements using contingent valuation: a comparison of techniques. *American Journal of Agricultural Economics* 70(1):21-28.

Briggs A.H., Goeree R., Blackhouse G. and O'Brien B.J. (2002). Probabilistic Analysis of Cost-effectiveness Models: Choosing Between Treatment Strategies for Gastro-esophageal Reflux Disease. *Medical Decision Making* 22: 290-308.

Briggs A., O'Brien B., and Blackhouse G. (2002). Thinking outside the box: recent advances in the analysis and presentation of Uncertainty in cost effectiveness studies. *Annual Review of Public Health* 23: 377-401.

Briggs A. and Fenn P. (1998). Confidence intervals or surfaces? Uncertainty on the cost effectiveness plane. *Health Economics* 7:724-740.

Brook C. (2002). Casemix funding for acute hospital care in Victoria, Australia. Victorian state government. Available online: casemix.health.vic.gov.au/about casemix updated 1 October 2002. September 16 2003 date last accessed.

Bryan S., Buxton M., Sheldon R. and Grant A. (1998). Magnetic resonance imaging for the investigation of knee injuries: an investigation of preference. *Health Economics* 7:595-604.

Cave M., Burningham D., Buxton M., Hanney S., Pollitt C., Scanlan M. and Schurmer M. (1993). *The valuation of changes in Quality in the Public Services*. HMSO, London.

Caves, D.W., Christenson, W.E. & Diewert, W.E. (1982). The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity. *Econometrica* 50:1393-1414.

Charnes A., Cooper R.C. and Rhodes E. (1978). Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research* 2: 429-444.

Chassin M.R. Hannan E.L. and DeBuono B.A. (1996). Benefits and hazards of reporting medical outcomes publicly. *New England Journal of Medicine* 334;6: 394-398.

Coelli T., Rao and Battese. (1998). *An Introduction to Efficiency and Productivity Analysis.* Kluwer Academic Publishers, 1998.

Coelli T. (1995). Recent developments in frontier modeling and efficiency measurement. *Australian Journal of Agricultural Economics* 39;3:219-245.

Collett D. (1994). *Modelling Survival Data in Medical Research*. London UK: Chapman and Hall, 1994.

Collopy BT. (2000). Clinical indicators in accreditation: an effective stimulus to improve patient care. *International Journal for Quality in Health Care* 12(3):211-216.

Commonwealth Department of Health and Aged Care. (2000). National Hospital Cost Data Collection 1998-99: final cost report (round 3), DHAC: Canberra, 2000. http://www.health.gov.au/casemix/costing/costw1.htm.

Cook W.D., Moez H. and Hans J. (2000). Multi-component efficiency measurement and shared inputs in data envelopment analysis: an application to sales and service performance in bank branches. *Journal of Productivity Analysis* 14; 3:209-224.

Coory M. and Gibberd R. (1998). New measures for reporting the magnitude of small-area variation in rates. *Statisitics in Medicine* 17: 2625-34.

Cox D.R. (1972). Regression models and life tables. *Journal Of The Royal Statistics Society Series B* 34:187-220.

Cropper, M.L. and W.E. Oates. (1992). Environmental Economics: A Survey. *Journal of Economic Literature* 30;2(June):675-740.

Culyer A.J., Van Doorsaler E. and Waggstaff A. (1992). Comment on utilisation as a measure of equity by Mooney Hall Donaldson and Gerard. *J Health Economics* 11(1):93-98.

Currie G.R., Donaldson C., O'Brien B., Stoddart G.L., Torrance G.W. and Drummond M.F. (2002). Willingness to Pay for What? A Note on Alternative Definitions of Health Care Program Benefits for Contingent Valuation Studies. *Medical Decision Making* 22:493-497.

Davies H., Nutley S. and Mannon R. (2000). Organisational culture and quality of health care. *Quality In Health Care* 9:111-119.

Diener A., O'Brien B. and Gafni A. (1998). Health care contingent valuation studies: a review and classification of the literature. *Health Economics* 7:313-326.

Donaldson C. and Gerard K. (1993). *Economics of Health Care Financing: The Visible Hand*. Macmillion, London.

Drummond M.F., Stoddard G.L. and Torrance G.W. (1987). *Methods for Economic Evaluation of Health Care Programmmes. Ist ed*. Oxford UK: Oxford University Press.

Drummond M.F., O'Brien B.J., Stoddard G.L. and Torrance G.W. (1997). *Methods for Economic Evaluation of Health Care Programmmes. 2nd ed*. Oxford UK: Oxford University Press.

Duckett S. (2000). The Australian Health Care System. Oxford University Press, Melbourne.

Duckett S. (1998). Case-mix funding for acute hospital inpatient services in Australia. *Medical Journal of Australia* 169:S17-S21.

Eckermann S. (1994). Hospital productivity: a contradiction in terms? In *Economics and Health 1994: Proceedings of the Sixteenth Australian Conference of Health Economists*. A.Harris (eds.). Australian Studies in Health Administration series. School of Health Administration, University of New South Wales, Sydney, 1994; 78: 167-181.

Eckermann S., Martin A., Stockler M. and Simes R.J. (2003). The benefits and costs of tamoxifen for breast cancer prevention. *Australian and New Zealand Journal of Public Health* 27;1:34-40.

Eckermann S. and Kirby A. (2003). "Cost Effectiveness Analysis: Uncertainty, Predictive Conditioning and Extrapolation Post Study" in Butler JRG and Quinn C (eds) Economics and Health: 2002 Proceedings of the Twenty-Fourth Australian Conference of Health Economists, AHES, Sydney (forthcoming).

Evans R.G. (1981). Incomplete Vertical Integration: The Distinctive Sructure of the Health Care Industry. In *Health, economics and Health Economics*, eds. J. Van der Gaag and M. Perlman, Amsterdam, North Holland: 329-354.

Evans R.G. (1984). *Strained Mercy: The Economics of Canadian Health Care*. Butterworths, Toronto.

Fahey P.P. and Gibberd R.W. (1995). Monitoring pulmonary embolisms in Australia - I. Variations between hospitals. *International Journal for Quality in Health Care* 7:373-380.

Färe R. and Grosskopf S. (2000). Outfoxing a paradox. *Economics Letters* 69:159-163.

Färe R., Grosskopf S. and Logan. (1983). The Relative Efficiency of Illinois Electric Utilities. *Resources and Energy* 5: 349-367.

Färe R., Grosskopf S. and Lovell C. (1985). *The Measurement of Efficiency of Production*. Boston. Kluwer-Nijhoff Publishing.

Färe R., Grosskopf S., Lovell C. and Parsuka C. (1989). Multi-product Productivity Comparison When Some Outputs are Undesirable: A Non Parametric Approach. *The Review of Economics and Statistics* 71; 1: 90-98.

Färe R., Grosskopf S., Lovell C. and Yaisawarang S. (1993). Derivation of Shadow Prices for Undesirable Outputs: A Distance Function Approach. *The Review of Economics and Statistics* 75;2:374-380.

Färe R., Grosskopf S. and Lovell C. (1994). Production Frontiers, Cambridge University Press, Cambridge.

Färe R., Grosskopf S., Lingdren B. and Roos P. (1994). Productivity Developments in Swedish Hospitals: a Malmquist Output Index Approach. In Charnes A, Cooper W, Lewin AY, Seiford LM (eds), *Data Envelopment Analysis: Theory, Methodology and Applications*, Kluwer Academic Publishers. (Originally presented at a conference of new uses of DEA in management and public policy, University of Texas , Ausin,TX, Sepetmber 27-29, 1989.)

Färe R, Grosskopf S and Roos P. (1995). Productivity and Quality Changes in Swedish Pharmacies. *The International Journal of Production Economics* 39:137-144.

Färe R., Grosskopf S. and Roos P. (1992). Productivity and quality changes in Swedish pharmacies 1980-1989, a nonparametric Malmquist approach, *Journal of Productivity Analysis* 3:85-101.

Färe R., Grosskopf S. and Roos P. (2002). Integrating consumer satisfaction into productivity indexes. In Fox K (eds), *Efficiency in the Public Sector*, Kluwer Academic Publishers.

Folland S.T. and Hofler R.A. (2001). How reliable are hospital efficiency estimates? Exploiting the dual to the homothetic production. *Health Economics* 10: 683-698.

Fox K.J. (1999). Efficiency at different levels of aggregation: public vs. private sector firms. *Economics Letters* 65:173-76.

Frazier A., Colditz G., Fuchs C. and Kuntz K. (2000). Cost-effectiveness of Screening for Colorectal Cancer in the General Population *Journal Of The American Medical Association* 284;15(18 October):1954-1961

Freedman D., Pisani R. and Purves R. (1980). Statistics. New York, Norton and Company, 1980.

Fried H.O., Schmidt S.S. and Yaisawarang S. (1999). Incorporating the operating environment into a non-parametric measure of technical efficiency. *Journal of Productivity Analysis* 12;3:249-267.

Gafni A. (1991). Willingness to pay as a measure of benefits. Relevant questions in the context of public decision making about health care programs. *Medical Care* 29:1246-52.

Gibberd R.W., Pathmeswaran A. and Burtenshaw K. (2000). Using Clinical Indicators to Identify Areas for Quality Improvement. *Journal of quality in clinical practice* 20:136-144.

Glasziou P., Simes R.J., Hall J. and Donaldson C. (1997). Design of a cost effectiveness study within a randomized trial: the LIPID trial for secondary prevention of IHD. *Controlled Clinical Trials* 18:464–76.

Glasziou P., Eckermann S., Mulray S., Simes R.J. Martin A., Kirby A. et al. (2002). Cholesterol lowering therapy with pravastatin in patients with average cholesterol levels and established ischaemic heart disease: is it cost effective? *Medical Journal of Australia* 177:420-426.

Goddard M., Mannion R. and Smith P. (2000). Enhancing performance in health care: a theory perspective of agency and the role of information. *Health Economics* 9:95-107.

Gold M.R., Siegel J.E., Russel L.B. et al. (1996). Cost Effectiveness in Health and Medicine. New York: Oxford University Press, 1996.

Goldberg P. (1995). Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry *Econometrica* 63;4:891-951.

Greene W. (1993). The Econometric Approach to Efficiency Analysis. In Fried, Lovell CAK and Schmidt SS (eds) *The Measurement of Productive Efficiency* Oxford University Press, New York 68-119, 1993.

Gregan T. and Bruce R. (1997). Technical efficiency in the hospitals of Victoria. In *A technique for measuring the efficiency of government service delivery*. Steering Committee for the Review of Commonwealth /State Service Provision Data Envelopment Analysis. Australian Government Publishing Service: Canberra, 1997.

Gronroos C. (1984). A Service Quality Model and its Market Implications. *European Journal of Marketing* 18;4:36-44.

Grosskopf S. and Valdmanis V. (1987). Measuring hospital performance: a non-parametric approach. *Journal Of Health Economics* 6:89-107.

Grossman M. (1972). On the concept of health capital and the demand for health. *Jounal of Political Economy* 80:223-255.

Hargreaves J. (2001). Reporting of adverse events in routinely collected data sets in Australia. Canberra, Australian Institute of Health and Welfare (Health Division Working Paper No.3), Canberra.

Harris J.E. (1977). The Internal organization of hospitals: some economic implications. *The Bell Journal of Economics* 8:467-482.

Haynes, K.E., S. Ratick, W.M. Bowen, and J. Cummings-Saxton. (1993). Environmental Decision Models: U.S. Experience and a New Approach to Pollution Management. *Environment International* 19:261-75.

Haynes, K.E., S. Ratick, and J. Cummings-Saxton. (1994). Toward a Pollution Abatement Monitoring Policy: Measurements, Model Mechanics, and Data Requirements. *The Environmental Professional* 16:292-303.

Henry C. (1974). Investment decisions under uncertainty: the irreversibility effect. *American Economic Review* December 1974: 1006-1012.

Heuchan A.M., Evans N., Henderson Smart D.J. et al. (2002). Perinatal risk factors for major intraventricular haemorrhage in the Australian and New Zealand Neonatal network, 1995 to 1997. *Archives of Diseases in Childhoo: Fetal and Neonatal Edition* 86:F86-90.

Hoch J.S., Briggs A. and Willan A. (2002). Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost effectiveness analysis. *Health Economics* 11:415-430.

Hollingsworth B., Dawson P.J. and Maniadakis N. (1999). Efficiency meaurement of health care: a review of non parametric methods and applications *Health Care Management Science* 2;3:161-172.

Holman C.D.J., Bass A.J., Rouse I.L., Hobbs M.S.T. (1999) Population-based linkage of health records in Western Australia: development of a health services research linked database. *Australian and New Zealand Journal of Public Health* 23 (5): 453-459.

Howley P.P. and Gibberd R.W. (2003). Using hierarchical models to analyse clinical indicators: a comparison of the gamma-Poisson and beta-binomial models. *International Journal for Quality in Health Care* 15;4:319-329.

Hunink M., Glasziou P., Seigel J., Weeks J., Pilskin J., Elstein A. and Weinstein M. (2001). *Decision Making in Health and Medicine*. Cambridge University Press.

Jacobs P. and Baladi J. (1996). Biases in cost measurement for economic studies in health care. *Health Economics* 5;6:525-529.

Jan S., Mooney M., Ryan M., Bruggerman K. andAlexander K. (2000). The use of conjoint analysis to elicit community preferences in public health research: a case study of hospital services in South Australia. *Australia and New Zealand Journal of Public Health* 24(1):64-70.

Johannesson M. (1993). The contingent valuation method – appraising the appraisers. *Health Economics* 2: 357-359.

Kahneman D. and Tversky A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica* 47(2): 263-291.

Kahnemann D., Knetsch J.L. and Thaler R.H. (1990). Experimental test of the endowment effect and the coarse theorem. *Journal of Political Economy* 98(6): 1325-1348.

Kelman C.W., Bass A.J. and Holman C.D.J. (2002). Research use of linked health data - a best practice protocol. *Australian and New Zealand Journal of Public Healt*h 26 (3): 251-255.

Kopp R.J. (1981). The Measurement of Productive Efficiency: a Reconsideration. *Quarterly Journal of Economics* 96:477-503.

Landon B., Lezzoni L., Ash A.S, Schwartz M., Daley J., Hughes J.S. and Mackerion Y.D. (1996). Judging hospitals by mortality rates: the case of CABG surgery. *Inquiry* 33(2):155-166.

Lee L., Eagar K. and Smith M. (1998). Subacute ad non acute case-mix in Australia. *Medical Journal of Australia* 169:S22-S25.

Long M.J., Chesney J.D. and Fleming S.T. (1990). A reassessment of hospital product and productivity changes over time. *Health Care Financing Review* 11;4:69-77.

Lovell C.A.K., Sakar A. and Sickles R.C. (1987). Output aggregation and the measurement of productive efficiency Science Working Paper Series no. 87-6 Department of Economics University of North Carolina.

Lovell C.A.K. (1993). Production Frontiers and productive efficiency. In *The measurement of productive effeiciency: Techniques and Applications*. Fried HO. Lovell AK Schmidt SS (eds). Oxford University Press. New York; 1993.

Lin D.Y., Freuer E.J., Etzioni R. and Wax Y. (1997). Estimating Medical Costs from Incomplete Follow-up Data. *Biometrics* 53: 419-434.

Magnussen J. (1996). Efficiency measurement and the operationalization of hospital production. *Health Services Research* 31:21-37.

Marshall M.N., Shekelle P.G., Leatherman S., Brook R.H. (2000). The public release of performance data: what do we expect to gain? A review of the evidence. *Journal of the American Medical Association* 283: 1866-1874.

Marschner I.C., Colquhoun D., Simes R.J., Glasziou P., Harris P. (2001). Long-Term Risk Stratification for Survivors of Acute Coronary Syndromes: Results From the Long_Term Intervention With Pravastatin in Ischeamic Disease (LIPID) Study. *Journal Of the American College Of Cardiology* 38: 56-63.

McGuire A., Henderson J. and Mooney G. (1988) The Economics of Health Care: *An introductory text.* Routledge, London.

McKee, M. (1997). Indicators of clinical performance: Problematic, but poor standards of care must be tackled [Editorials]. *British Medical Journal* 315;7101(19 July):142.

Meeusen W., Van den Broeck J. (1977). Efficiency estimation from Cobb-Douglas Production Functions with composed error. *International Economic Review* 18:435-444.

Ministry of Health of Ontario. (1994). *Ontario Guidleines for Economic Analysis of Pharmaceutical Products.* Ministry of Health of Ontario: Toronto, 1994.

Mobley L.R. and Magnussen J. (1998). An international comparison of hospital efficiency: does institutional environment matter? *Applied Economics* 30;8:1089-1100.

Mooney G.H., Russell E.M. and Weir R.D. (1986). *Choices for Health Care*. Macmillan, London.

Mooney G.H., Gerard K., Donaldson C. and Farrar S. (1992) Priority Setting in Purchasing: some practical guidelines. NAHAT, Birmingham.

Mooney G.H. (1994a). *Key Issues in Health Economics*. Harvester Wheatsheaf: New York, 1994.

Mooney G.H. (1994b).What else do we want form our health services? *Social Science and Medicine* 39:151-154.

Mooney G.H. and Lange, M. (1993). Ante-Natal Screening: what constitutes benefit? *Social Science and Medicine* 37:873-878.

Morey R.C., Fine D.J., Loree S.W., Retlaff-Roberts D.L. and Tsubakitani S. (1992). The tradeoff between hospital costs and quality of care. An exploratory empirical analysis. *Medical Care* 30:677-698.

National Health Service. (2002). Performance Indicators. Department of Health. http://www.doh.gov.uk/nhsperformanceindicators/2002/trust.html: February 2002.

National Health Care Purchasing Institute. (2002). Ensuring quality providers: a purchaser's toolkit for using incentives. http://www.nhcpi.net/pdf/models.pdf National Health Care Purchasing Institute: Washington.

National Health Performance Committee. (2000). *Fourth national report on health sector performance indicators*. NSW Health, Sydney.

National Institute for Clinical Excellence. (2001). *Technical Guidance for Manufacturers and Sponsors on Making a Submission to a Technology Appraisal*. National Institute for Clinical Excellence: London, 2001.

National Oceanic and Atmospheric Administration. (1993). Report of the NOAA panel on contingent valuation. *Federal Register* 58: 4607-14.

Newhouse J. (1970) Towards a Theory of Non-Profit Insitutions: An Economic Model of a Hospital. *Amercian Economc Review* 60(1):64-74.

Newhouse J. (1994). Frontier Estimation: How useful a tool for health economics? *Journal Of Health Economics* 13:317-322.

O'Brien B.J., Gersten K., Willan A.R., Faulkner L.A. (2002). Is there a kink in consumers' threshold value for cost effectiveness in health care? *Health Economics* 11:175-180.

O'Brien B.J. and Gafni A. (1996). When do dollars make sense? Towards a conceptual framework for contingent valuation studies in health care. *Medical Decision Making* 16:288-99.

O'Brien B.J., Goeree R., Gafni A., Torrance G., Pauly M., Erder H., Rusthoven J. Weeks J., Cahill M. and LaMont B. (1998). Assessing the Value of a New Pharmaceutical: A Feasibility Study of Contingent Valuation in Managed Care. *Medical Care*. 36;3(March):370-384.

Opaluch, J. Swallow, S., Weaver T., Wessels C. and Wichens C. (1993). Evaluating impacts form noxious facilities: using public preferences in current siting mechanisms. *Journal of Environmental Eonomics and Management* 24:41-59.

Pekurinen M., Sintonen H., Pitaken E., Alander V. and Coyle D. (1991). Hospital productivity in Finland: further analysis. *The Finnnish Journal of Business Economics*. 1991; Special edition 1, Helsinki, Finland.

Pettiti D. (1994). Meta-Analayis, Decision Analaysis and Cost-Effectiveness Analysis: Methods for quantitative synthesis in Medicine. Oxford University Press.

Pindyck R.S. (1988). Irreversible investment, capacity choice and the value of the firm. *American Economic Review* December 1988: 969-985.

Pindyck R.S. (1991). Irreversibility, Uncertainty and Investment. *Journal of Economic Literature* September 1991: 1110-1148.

Pittman, R.W. (1981). Issues in Pollution Control: Interplant Cost Differences and Economies of Scale. *Land Economics* 57;February:1-17.

Propper C. (1990) Contingent valuation of time spent on the NHS waiting lists. *The Economic Journal* 1990;100:193-9.

Propper C. (1995). The disutility of time spent on the United Kingdoms waiting lists. *The Journal of Human Resources* 1995; 30:677-700.

Puig-Junoy J. (1998).Technical efficiency in the clinical management of critically ill patients. *Health Economics* 7: 263-277.

Rheinhard S., Lovell K., Thjissen G. (1999) Econometric Estimation of Technical and Environmental Efficiency: An Application to Dutch Dairy Farms. *American Journal of Agricultural Economics* 81;1:44-60.

Rice N. and Smith P.C. (2001). Capitation and Risk Adjustment in Health Care Financing. *Millbank Quarterly* 79(1):91.

Roberts C.M., Barnes S., Lowe D. and Pearson M.G. (2003). Evidence for a link between mortality in acute COPD and hospital type and resources. *Thorax*. 58(11):947-949.

Roos P. (2002). Measuring Output of Hospital Services. In Fox K (eds), *Efficiency in the Public Sector*. Chapter 9:249-271. Kluwer Academic Publishers.

Ryan M. and Hughes J. (1997). Using conjoint analysis to assess women's preferences for miscarriage management. *Health Economics* 6:261-273.

Ryan M., McIntosh E. and Shakley P. (1998a). Methodological issues in the application of conjoint analysis in health care. *Health Economics* 7:373-378.

Ryan M., McIntosh E. and Shakley P. (1998b). Using conjoint analysis to consumer preferences in primary care; an application the patient health card. *Health Expectations* 1998; 1:117-129.

Ryan M. (1999). Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilization. *Social Science and Medicine* 48:535-546.

Salkeld G., Ryan M. and Short L. (2000). The veil of ignorance. Do consumers prefer what they know best? *Health Economics* 9:267-270.

Seiford, L.M. and Thrall R.M. (1990). Recent Developments in DEA: The Mathematical Programming approach to Frontier Analysis. *Journal of Econometrics* 46;1/2:7-38.

Sen, A. (1993) Capability and Well Being. In *The Quality of Life*, Nussbaum MC and Sen A (eds). Oxford: Oxford University Press.

Shephard R.W. (1953) Theory of Cost and Production Functions. 1970 edition, Princeton, NJ: Princeton University Press.

Shephard R.W. (1974). *Indirect Production Functions*. Verlag Anton Hain, Meisenheim am Glan.

Sherman H.D. (1984). Hospital efficiency measurement and evaluation: empirical test of a new technique. *Medical Care* 22;10:922-938.

Simon, H. (1957). Models of Man. Wiley, New York.

Simpson J.M., Evan N., Gibberd R.W., Heuchean A. and Henderson-Smart D. (2003). Analysing differences in clinical outcomes between hospitals. *Quality and Safety in Health Care* 12:257-263.

Smith P. (2002). Measuring health system performance. *European Journal of Health Economics* 3:145-148.

Smith P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration* 18:277-310.

Stigler GJ. (1976). The Xistence of X-Efficiency. *American Economic Review* 66(1):213-216.

Stinnett A.A. and Mullahy J. (1998). Net Health Benefits: a new framework for the analysis of uncertainty in cost effectiveness analysis. *Medical Decision Making* 18 (2 suppl): S65-S80.

Stinnett, AA. and Paltiel, D. (1997). Estimating C/E ratios under second order uncertainty: the mean ratio versus the ratio of the means. *Medical Decision Making* 17, 483-489.

Street A. (2003). How much confidence should we place in efficiency estimates? *Health Economics* 12; 895-907.

Thanassoulis E., Boussofiane A. and Dyson R.G. (1995). Exploring output quality targets in the provision of perinatal care in England using data envelopment analysis. *European Journal of Operational Research* 80:588-607.

The Long-Term Intervention with Pravastatin in Ischaemic Disease

(LIPID) Study Group. (1998). Prevention of cardiovascular events and death with pravastatin in patients in patients with coronary heart disease and a broad range of initial cholesterol levels. *New England Journal of Medicine* 339:1349–57.

Thomson R.G., Singleton F.D., Thrall R.M. and Smith B.A. (1986). Comparative site evaluations for locating a high energy physics lab in Texas. *Interfaces* 16:35-49.

Thomson R.G., Langemeier L.N., Lee C.T. and Thrall R.M. (1990). The role of multiplier bounds in efficiency analysis with application to Kansas farming. *Journal of Econometrics* 46: 93-108.

Thomson R.G., Lee C.T. and Thrall R.M. (1992). DEA/AR efficiency of U.S. independent oil/gas producers over time. *Computers and Operations Research* 19: 377-391.

Tversky A. and Kahnemann D. (1981). The framing of decisions and the psychology of choice. *Science* 211:453-458.

Valdmanis V. (1992). Sensitivity analysis for DEA models: an empirical example using public vs NFP hospitals. *Journal of public economics* 48(2):185-205.

Van der Pol M. and Cairns J. (1998). Establishing preferences for blood transfusion support: an application of conjoint analysis. *Journal of health service research and policy* 3:70-6.

Vick S. and Scott A. (1998). What makes a perfect agent? A pilot study of patient preferences in the doctor-patient relationship. *Journal Of Health Economics* 17:587-606.

Wagstaff A. (1989). Estimating efficiency in the hospital sector; a comparison of three statistical cost frontier models. *Applied Econometrics* 21(5):659-672.

Wardman M. (1988). A comparison of revealed preference and stated preference models. *Journal of Transport Economics and Policy* 22:71-91.

Webster R,, Kennedy S. and Johnson L. (1998). *Comparing techniques for measuring the efficiency and productivity of Australian private hospitals.* ABS working papers in econometrics and applied statistics 98/3. Cat. no. 1351.0. Australian Bureau of Statistics: Canberra 1998.

Weinstein M.C. and Fineberg H.V. (1980). *Clinical Decision Analaysis*. W.B. Saunders: Philadelphia.

Weinstein M.C. (1995). From cost-effectiveness ratios to resource allocation: where to draw the line? In: Sloan FA ed. Valuing Health Care. Cambridge, UK: Cambridge University Press: 77-98.

Weisbrod B. (1991). The Quality Care Quadrilemma: An Essay on Technical Change, Insurance, Quality of Care and Cost Containment. *Journal of Economic Literature* 1991 Vol.XXIX: 523-552.

Willan A.R., O'Brien B.J. and Leyva R.A. (2001). Cost effectiveness analysis when the WTA is greater than the WTP. *Statistics in Medicine* 20:3251-3259.

Wilson R.M., Runciman W.B., Gibberd R.W. (1995). The Quality in Australian Heatlh Care Study. Medical Journal of Australia 163:458-471.

Wolfson M. and Alvarez R. (2002). Towards integrated and coherent health information systems for performance monitoring: the Canadian experience. In Smith P. (ed.) Measuring up: improving health systems performance in OECD countries. OECD: Paris, 2002.

World Heath Organisation. (1948). Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22

June 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.

Zuckerman S., Hadley J. and Lezzoni L. (1994). Measuring Hospital Efficiency with Frontier Cost Functions. *Journal Of Health Economics* 13:255-280.