THE UNIVERSITY *of York*

CENTRE FOR HEALTH ECONOMICS

Measuring performance: An examination of composite performance indicators

Rowena Jacobs
Peter Smith
Maria Goddard

CHE Technical Paper Series 29

CENTRE FOR HEALTH ECONOMICS TECHNICAL PAPER SERIES

The Centre for Health Economics has a well established Discussion Paper series which was originally conceived as a means of circulating ideas for discussion and debate amongst a wide readership that included health economists as well as those working within the NHS and pharmaceutical industry.

The introduction of a Technical Paper Series offers a further means by which the Centre's research can be disseminated. The Technical Paper Series publishes papers that are likely to be of specific interest to a relatively specialist audience, for example papers dealing with complex issues that assume high levels of prior knowledge, or those that make extensive use of sophisticated mathematical or statistical techniques.

The content and its format are entirely the responsibility of the author, and papers published in the Technical Paper series are not subject to peer-review or editorial control, unlike those produced in the Discussion Paper series. Offers of further papers, and requests for information should be directed to Frances Sharp in the Publications Office, Centre for Health Economics, University of York.

March 2004.

PLEASE DO NOT QUOTE WITHOUT THE AUTHORS' PERMISSION

Measuring performance: An examination of composite performance indicators

A report for the Department of Health

Rowena Jacobs, Peter Smith, Maria Goddard

Centre for Health Economics
University of York
York, Y010 5DD
United Kingdom
Tel: +44-1904-321425
Fax: +44-1904-321454
Email: rj3@york.ac.uk

March 2004

Abstract

It is widely acknowledged that the performance of health care systems is multidimensional, implying that measurement of performance requires a number of different scales to capture each element. Given the increasing interest worldwide in combining disparate indicators of healthcare performance into a single composite measure, this report examines the economic and technical issues involved in the creation of composite indicators. There are many arguments for and against the use of composite indicators and while they are increasingly being used as a political tool in a number of different policy areas, they can potentially suffer from a number of methodological difficulties. The publication of composite performance indicators might generate both positive and negative behavioural responses depending on the incentives which they produce. A number of different examples are given of the creation and use of composite indices in health care and in other sectors, both in the UK and internationally.

One example, is the introduction of the annual "star ratings" of NHS hospitals in the UK aimed to create a rounded picture of performance by combining several dimensions of hospital output. This report uses this example of hospital data for English acute hospitals to examine the methodological challenges of creating composite measures. The creation of a composite comprises a number of important steps, each of which requires careful judgement. These include the specification of the choice of indicators, the transformation of measured performance on individual indicators, combining the indicators using some decision rules and the specification of a set of weights that reflect societal preferences for health system outputs. The report examines these issues by developing a composite index and using Monte Carlo simulations to examine the robustness of performance judgements to these different technical choices. Model uncertainty is explored by changing assumptions about random variation in the indicators and then examining the impact this has on hospital rankings.

The analysis suggests that the technical choices that have to be made in the construction of the composite can have a significant impact on the resulting score. In particular, changing the weights, thresholds and decision rules of combining individual performance indicators materially affects the score and rank correlations of hospitals. Technical and analytical issues in the design of composite indicators clearly have important policy implications. This report highlights the issues which need to be taken into account in the construction of robust composite indicators so that they can be designed in ways which will minimise the potential for producing misleading performance information which may fail to deliver the expected improvements or even induce unwanted side-effects.

Keywords: performance measurement, performance indicators, composite indicators

Executive summary

1. A composite indicator is an aggregated index comprising individual performance indicators. Composite indicators are useful as a tool for conveying summary performance information and signalling policy priorities and are used widely in health care and other sectors, nationally and internationally.

2. Composite performance indicators have a number of advantages, such as focusing attention on important policy issues, offering a more rounded assessment of performance and presenting the 'big picture' in a way in which the public can understand.

3. However, the construction of composite measures is not straightforward and the associated methodological challenges raise a series of technical and economic issues that, if not addressed adequately, can create the potential for composite measures to be misinterpreted or manipulated. Moreover, the dangers are not restricted to technical issues as the use and publication of composite performance measures can generate both positive and negative behavioural responses, so careful consideration needs to be given to their creation and subsequent use.

4. This report uses the data created for the NHS Star Ratings for acute hospitals in order to demonstrate the importance of the methodological decisions taken at each stage in the construction of composite indicators, illustrating the impact of various choices on the final measure and on the potential incentive effect of the indicators. Rather than making a critique of the specific Star Rating system, it is used merely as a method of examining empirically the issues involved in the construction of composite measures.

5. The report first describes the use of composite indicators of performance in health care and other sectors in the UK and internationally.

6. Each step in the creation of a composite measure is then outlined in detail, with the main methodological issues highlighted through the use of examples of composite measures created elsewhere. The steps involved in constructing the composite include:

   - choosing the entities to be assessed
   - choosing the organisational objectives to be encompassed in the composite
   - choosing the indicators to be included
   - transforming measured performance on individual indicators
   - combining the individual measures using addition or some other decision rules
   - specifying an appropriate set of weights
   - adjusting for environmental or other uncontrollable influences on performance
   - adjusting for variations in expenditure if a measure of efficiency is required
   - using sensitivity analysis to test the robustness of the composite to the various methodological choices

7. The empirical analysis using Star Ratings data from 2001/2 has two components. First the underlying nature of the data is explored, including the distributions of the underlying indicators, how they have been transformed, the correlations between the indicators and a factor analysis. Second, the dataset is used to construct a new composite index through a simulation exercise. Each of the steps of constructing a new composite is simulated and the robustness of the subsequent rankings of individual hospital trust units are examined.

8. The exploratory data analysis raised a number of important technical issues including: significant skewness in some variables, especially the key targets, with little variation in measured performance; issues around the level of data aggregation of some indicators; the arbitrary nature of the thresholds chosen for transforming variables; and high levels of correlations between some indicators. The results suggest the following sets of independent factors emerge from the data: inpatient satisfaction, waiting times, cancelled operations, readmission rates, cancer waits, financial balance and the CHI review.

9. Monte Carlo simulations were then used to develop a new composite index based on ten indicators chosen from the exploratory data analysis. Uncertainty intervals were created around the composite and the results suggest that the intervals overlap over the entire range of the composite. This casts doubt on whether one hospital is performing better than another and leads to concerns over the ranking of hospitals on the basis of the composite.

10. The empirical analysis then proceeds to use the new composite (and the uncertainty intervals) to examine a number of important issues, namely: decomposing the variation on the performance indicators (to separate out real differences in performance from those due to random variation, measurement error, sampling error or simply natural variation); introducing different weighting systems; transforming the indicators to a categorical scale; and introducing decision rules to construct the composite.

11. The analysis suggests that technical and analytical choices made in constructing the composite measures can have a significant impact on the resulting score. In particular, changing the weights, thresholds and decision rules for combining individual indicators materially affects the score and the rankings of hospitals.

12. This research indicates that methodological issues in the design of composite indicators are of interest not merely from a technical perspective, but because they also have important policy implications. The report highlights the issues which need to be taken into account in the construction of robust composite indicators so they can be designed in ways which will minimise the potential for producing misleading performance information. If such issues are not addressed, composite measures may fail to deliver the expected improvements in performance or may even induce unwanted side-effects.

Table of contents

List of tables

List of figures

1. Introduction

A composite indicator is an aggregated index comprising individual performance indicators. It is an index of relative attainment since it reflects the relative values of the underlying individual performance indicators. Composite indicators are increasingly being used to measure the performance of, and also to rank, organisations and institutions in economic, social and policy areas (Freudenberg, 2003). Composite indicators integrate a large amount of information in a format that is easily understood and are therefore a valuable tool for conveying a summary assessment of performance in priority areas. In the English NHS, composite indictors in the form of Star Ratings have been accorded a high profile as a means of securing performance improvement across the health care sector (Department of Health, 2001).

Despite the proliferation of composite indicators across various sectors, the construction of a composite indicator is not straightforward methodologically. This leaves it open to misinterpretation and potential manipulation. Questions of the accuracy, reliability and appropriateness of such indices, need to be addressed if major policy, financial and social decisions hinge on the results of composite indicators.

This report examines the advantages and disadvantages of constructing a composite indicator and explores the technical and economic issues arising from the methodological choices made at each step in the construction of a composite. These include choosing the indicators to be used, transforming or standardising the indicators, applying a system of weights to the indicators and then combining them to form the new composite. Data from the Star Rating system for acute hospitals in England is used to explore empirically the methodological issues involved in each step of the process of constructing the composite indicator.

The NHS Star Ratings were the first published composite index for acute hospital trusts (Department of Health, 2001). Using a four-point scale, the ratings assessed the performance of acute general trusts in 2000/01 on certain ‘key targets’ with the additional criterion that the trust did not receive a critical clinical governance report from the Commission for Health Improvement (CHI). The targets comprise various waiting time indicators, hospital cleanliness, cancelled operations, the financial state of the trust and the demonstration of a commitment to improve working conditions. Three sets of subsidiary indicators were also assessed, including those with a clinical focus (such as emergency re-admission rates); those with a patient focus (such as the resolution of written complaints) and those with a staff focus (such as junior doctors’ hours). Trusts with a ‘balanced scorecard’ – good performance in all four areas – were then awarded two or three stars respectively.

The Star Rating results receive wide publicity and various incentives are also attached to performance as measured by the rating. For example, trusts rated as having three stars benefited from ‘earned autonomy’ which awards them greater operational freedoms and less intensive monitoring and are able to apply for foundation status, gaining even greater independence from the centre. Over the last two years, NHS Star Ratings have evolved to include additional and more sophisticated indicators and to cover other types of organisation such as Primary Care Trusts (PCTs).

Rather than commenting specifically on the appropriateness of the Star Rating system as a means of capturing performance in the NHS, this report merely uses the ratings data to explore the generic issues involved in constructing a composite index. The report examines the methodological challenges by constructing a composite index and using Monte Carlo simulations to examine the robustness of performance judgements to different technical choices made at each stage of the process. Uncertainty is explored by changing the assumptions made about the degree of random variation in the underlying indicators and examining the subsequent impact on hospital rankings.

The report first describes briefly the use of composite indicators in health care and other sectors, highlighting the techniques employed in their construction and drawing out the potential methodological pitfalls. The empirical work using the Star Ratings data as described above is then presented, before concluding with some analytical and policy considerations, as well as avenues for future research.

2. Arguments for and against the use of composite indicators

Composite indicators have a high profile in the media and potentially have an important role alongside the publication of individual performance indicators.

Some of the arguments for developing composite indicators include the following (Smith, 2002):

1. They place performance at the centre of the policy arena
2. They can offer a rounded assessment of performance
3. They enable judgements to be made on system efficiency
4. They facilitate communication with the public and promote accountability
5. They indicate which organisations represent beacons of best performance
6. They indicate priority organisations for improvement efforts
7. They may stimulate the search for better analytical methods and better quality data
8. They present the ‘big picture’ and can be easier to interpret than trying to find a trend in many separate indicators.

On the other hand, composite indicators may lead to a number of dysfunctional consequences and there are a number of arguments against their use including the following (Smith, 2002):

1. By aggregating individual performance measures, composites may disguise serious failings in some parts of the system
2. As measures become aggregated it becomes more difficult to determine the source of poor performance and where to focus remedial action
3. The individual performance measures used in the composite are often contentious
4. A composite that is comprehensive in coverage may have to rely on poor quality data in some dimensions
5. A composite that ignores some performance measures because they are difficult to measure may distort behaviour in undesirable ways
6. The composite measure will depend crucially on the weights attached to each performance dimension. However, the methodology by which weights are

elicited and decisions on whose preferences they should reflect are unlikely to be straightforward.

3. Responses to performance information

An important consideration in the creation of composite performance indicators, is the response which their publication will generate. The belief underpinning the publication of composite and other performance data is that it will improve overall performance by encouraging consumers, or purchasers on behalf of consumers, to choose high quality providers. Providers can use comparative performance data to assess their position relative to others. It is also argued that the performance data can be used as a tool to regulate and ensure accountability (Marshall *et al*, 2000). In the US, there is an increasing tendency for performance information to be collected and disseminated and summarised in the form of report cards for various provider organisations (Lubalin and Harris-Kojetin, 1999).

The impact of the publication of performance data on behaviour is however uncertain, with little evidence that consumers or health care professionals trust and use it, despite the fact that consumers say this is the sort of information they require (Sheldon *et al*, 2001; Dranove *et al*, 2002; Schneider and Lieberman, 2001). Various reasons have been put forward for this apparent contradiction (Bernstein and Gauthier, 1999), including a lack of understanding, a lack of trust in the source of the information, difficulty in processing conflicting information, or the information does not accord with prior beliefs. In fact, consumers appear to continue to use hospitals with high mortality rates since they appear to rate anecdotal evidence from family and friends more highly than risk-adjusted mortality data (Marshall *et al*, 2000), even though friends and family may not be the best judges of the more technical characteristics of provision (McLaughlin, 1999). Furthermore, the public may judge quality of service not only in terms of clinical effectiveness, but also accessibility, ease and pleasantness of use, the staff, and the information provided. However, there is some evidence that the younger and better educated appear to use performance information now to a greater extent than other groups (Marshall *et al*, 2002).

When faced with lots of performance information, consumers will need to weigh up the evidence and make trade-offs between different performance dimensions, thus increasing their processing burden. Some consumers may end up basing decisions on a single performance dimension simply because it is the most clear but not necessarily the most important to them. This underscores the potential advantage of presenting performance information to consumers in the form of a composite indicator.

Provider organisations appear to be more responsive to performance data than consumers and more likely to use the information for benchmarking or internal monitoring of performance. Evidence suggests that publication can be a catalyst for provider behaviour change which can contribute to observed improvement in processes and outcomes of care. A few studies have shown that public disclosure of performance data has resulted in a significant reduction in risk-adjusted mortality (Hannan *et al*, 1994). Sometimes publication of data can also result in checking and auditing of the quality of data rather than direct action to improve delivery of service (Mannion & Goddard, 2001). Providers appear to be sensitive to their public image.

However the publication of performance data can also have a number of adverse consequences such as gaming and the ‘ratchet effect’ if performance is monitored relative to own past performance. Thus providers may have the incentive to purposefully under-perform in order to ensure less stringent future performance targets. This may lead to an inappropriate focus on what is being measured in the composite. The publication of performance data may also lead to data manipulation, mis-reporting or creative accounting through adverse patient reports unaccountably being lost or excluded, risk characteristics being exaggerated, or activity figures artificially being boosted (Nutley & Smith, 1998; Goddard & Smith, 2001). It may result in problems finding consultants to undertake procedures on high-risk patients (Schneider & Epstein, 1996) if adequate risk-adjustments are not made to the data. It may also have a negative impact on professional morale (Marshall *et al*, 2000).

Careful consideration therefore needs to be given to the intended use of the performance indicators, the composite and associated targets. If they are largely ignored by the public and purchasers for decision-making purposes, then their publication may simply send reassuring signals to the electorate that action is being taken on managing performance (Street, 2002). If they are being used purely internally by providers as a source of information for monitoring performance, supplemented by other detailed local data, then proper contextualisation of the information and explanations of the limitations of the performance indicators and the composite may not be necessary. If however the data is used largely for the construction of league tables and such performance rankings might affect budgets, bonuses and job security, influence morale and recruitment, or even lead to a loss of contracts, then its publication, accuracy, and timeliness will need to be carefully scrutinised (Smith, 1995). Reported differences in performance which are statistically insignificant, may also then be entirely inappropriate (Nutley & Smith, 1998). The arguments outlined above apply to all types of performance information, not just composite indicators. However, many of the issues may assume more importance when composite indices are considered as there is a tendency for them to receive greater attention and to be viewed as an easy way of assessing relative performance.

4. International examples of the development and use of composite indicators

In this section the development and use of composite indicators, in health care and other public sector, in the UK and internationally, are described briefly to provide some general background and an overview of the methodologies employed.

4.1. Health care

4.1.1. United States Medicare

In 2000 and 2003 Jencks *et al* produced a series of 22 quality indicators of the care delivered to Medicare beneficiaries (primarily in fee-for-service) and constructed state-level composite indices from these indicators. They examined these state-level indicators for the periods of 1998/99 (baseline) and 2000/01 (follow-up) and examined changes in performance across the range of indicators. The quality indicators were abstracted from state-wide random samples of medical records for inpatient fee-for-service care (16 indicators) and from Medicare beneficiary surveys or Medicare claims for outpatient care (6 indicators).

Indicators were chosen for 6 clinical areas, namely: acute myocardial infarction (6 indicators), heart failure (2 indicators), stroke (3 indicators), pneumonia (7 indicators), breast cancer (1 indicator), and diabetes (3 indicators). The choice of indicators tended to over-represent inpatient and preventive services and under-represent ambulatory care and interventional procedures. The indicators were also not risk-adjusted and hence focused on process measures rather than outcomes. Clinical topics were selected according to five criteria:

- The disease is a major source of morbidity or mortality;
- Certain processes of care are known to improve outcomes;
- Measurement of these processes is feasible;
- There is substantial scope for improvement in performance;
- Managerial intervention can potentially improve performance.

Reliability assessments were carried out at each abstraction centre on a monthly random sample of 30 cases taken from abstracts in the previous month. Reliability was calculated as the percentage of agreement on all abstraction data elements between two blinded independent abstractors at different centres. Median inter-rater reliability was 90 percent, though it ranged from 80 to 95 percent.

In the first study (Jencks *et al*, 2000), each of the 52 states is ranked on each of the measures, thus the percentage score is transformed to an ordinal scale ranging from 1 to 52. A composite performance measure was produced by computing each state's average rank.

In the follow-up study (Jencks *et al*, 2003), absolute improvement was also calculated on each indicator (defined as the percentage improvement from baseline to follow-up) and relative improvement, or the reduction in error rate, (defined as the absolute improvement divided by the difference between baseline and perfect performance (100%)).

To summarize the overall changes in performance at the state level, they calculated the median amount of absolute and relative improvement across the set of indicators in the state. They also calculated the rank of each state on each quality indicator based on the 2000/01 performance and the rank on each quality indicator based on relative improvement. They then calculated the average rank for each state across the 22 indicators and league-tabled them according to their average rank based on 2000/01 performance as well as relative improvement. They found that a state's average rank on the 22 indicators was highly stable over time with a correlation of 0.93 between the two periods.

The better performing states appeared to be concentrated geographically in the northern and less populated regions (for both periods) while the geographic patterns of relative improvement by state were more patchy. While the report showed overall improvement across 20 of the 22 indicators and a median relative improvement of 12.8 percent (in the median state), the cross-sectional data used could not provide information about the source of the quality improvement and the role of quality improvement efforts in that.

There were some concerns about the choice of indicators (process measures) which the authors acknowledged as well as the data reliability of some of the indicators (Smith, 2002). Furthermore, since all indicators were given equal weight, the use of eight indicators for AMI would give that clinical area more of a contribution in the composite than say breast cancer for which there is only one indicator.

The use of the league table ranking as the basis for the composite also implicitly assumes that identical differences in ranking are equally important, regardless of where in the league table they occur. The incentive is therefore for states to concentrate on activities where they can more readily secure a movement up the league table, rather than those that offer the most potential health gain.

4.1.2. Canadian regional health care

Macleans magazine (MM) is a major mass circulation Canadian magazine that publishes an annual "Health Report" in which they rank Canadian regions according to their healthcare. This is done on the basis of data published by the Canadian Institute for Health Information on a series of annual reports as well as a series of health indicators for the 63 largest regions (covering 90 percent of the population) (Canadian Institute for Health Information, 2001a; Canadian Institute for Health Information, 2001b).

The third *Macleans* report in 2001 used 15 health care performance indicators grouped into six categories:

1. *Outcomes*: 1) Life expectancy at birth, 2) Heart attack survival;
2. *Prenatal care*: 1) Proportion low birthweight babies under 2500g, 2) Percentage of babies born by Caesarean section, 3) Percentage of vaginal births after Caesarean section;
3. *Community health*: 1) Hip fractures, 2) Pneumonia and flu hospitalisation of persons over 64;
4. *Elderly services*: 1) Hip replacements, 2) Knee replacements;
5. *Efficiency*: 1) Possible outpatients – hospitalisations for conditions not requiring admission, 2) Early discharge – variation from expected length of stay, 3) Preventable admissions - hospitalisations for conditions considered preventable by appropriate ambulatory care; and
6. *Resources*: 1) Physicians per 100,000; 2) Specialists per 100,000; 3) Local services – percentage of hospitalisations generated by local residents.

The MM report evaluated 54 regions with populations over 125,000, classified as either: communities with medical schools, other major communities, or largely rural communities.

MM rescaled each of the 15 indicators to have a mean of 80 and a standard deviation of 10 (with a higher score implying better performance). Where there was missing data, scores were inferred from performance on non-missing data. Within each of the six categories, the scores on the performance indicators were combined using weights ‘based on expert opinion’. The six categories were then combined using the following weights: outcomes 0.2; prenatal care 0.2; community health 0.2; elderly services 0.1; efficiencies 0.2; resources 0.1. This sequential approach to assigning weights, first to performance indicators within categories and then to categories, allows a more careful treatment of priorities, albeit that the weighting scheme is very rudimentary and the preferences of the ‘experts’ may not necessarily reflect those of the public.

The composite performance scores ranged from 89.5 in North/West Vancouver, British Columbia to 73.4 in North Bay/Huntsville, Ontario in 2001.

MM acknowledge that there is very little variation in some of the indicators, however they discuss these variations as if clear differences exist which reflect real differences in health services. MM includes no psychiatric performance indicators which is a major aspect of health care services (Page & Cramer, 2001). Moreover the inclusion of efficiency indices leads efficiency to be treated not as the extent to which objectives are secured in relation to expenditure, but as simply another objective that contributes to the concept of performance (Smith, 2002).

4.1.3. British health authorities

In 2000, the UK television broadcaster Channel 4 commissioned researchers at the Kings Fund to explore the public’s relative preferences or health care priorities (Appleby & Mulligan, 2000). They produced a ranked list of English and Welsh health authorities and Scottish health boards according to a composite indicator based on selected aspects of performance, designed to reflect the relative weight attached by the public to these measures of NHS performance. Researchers were limited on the number of indicators with which they could feasibly survey the public and thus restricted their choice to six indicators chosen from readily available data produced by the NHS Executive (the High Level Performance Indicator set from the Performance Assessment Framework (PAF)):

1. Number of deaths from cancer (per 100,000)
2. Number of deaths from heart disease (per 100,000)
3. Number of people on hospital waiting lists (per 1,000)
4. Percentage of people on waiting lists for more than 12 months
5. Number of hip operations (per 100,000)
6. Number of deaths from ‘avoidable’ diseases (tuberculosis, asthma etc. for which there are effective clinical interventions that would prevent death) (per 100,000)

The focus of the study was to attach weights to each indicator based on public preferences. A polling organisation, MORI, surveyed 2000 people across England, Scotland and Wales to obtain their preferences. Three methods were used for eliciting preferences, namely ranking from most to least desired indicator, budget-pie, where respondents were asked to allocate a ‘budget’ of 60 chips between the six performance indicators and conjoint analysis, a more complicated technique asking respondents to choose between different mixes of options. Statistical analysis was then used to extract the average value placed on each indicator by the respondents. All three methods produced very similar overall rankings of health authorities and health boards based on the final composite indicator.

The weights for each indicator were taken from the budget-pie method and are shown in Table 1 with the weighting for hip operations taking a negative value indicating the "more is better" nature of the value. The weights were then multiplied by the actual performance values for each health authority and health board and then summed to produce a composite score, weighted by public preferences.

Table 1: Weights for six performance indicators based on 'budget-pie' survey

Indicator	Number of chips	Weight
Reducing deaths from cancer	16	1.00
Reducing deaths from heart disease	12	0.75
Reducing total number of people on hospital waiting lists	10	0.63
Reducing number of people waiting more than 12 months	9	0.56
Reducing deaths from 'avoidable' diseases	5	0.50
Increasing number of hip operations	8	-0.31

Source: Appleby & Mulligan (2000)

The researchers were concerned that some of the raw performance indicators had skewed distributions and were not all measured on the same scale. The transformed data (taking the square root) were standardised by taking the difference between the transformed values for each performance indicator for each health authority and the average value for the UK and expressing this as a proportion of the standard deviation for each indicator (resulting in a z score). The rankings generated from these transformed and standardised indicators differed from the original ranks with the average change in ranking being 14 places and a 0.81 rank correlation. The researchers then explored the extent to which socio-economic factors explained the variation in performance by controlling for socio-economic characteristics of the population. They controlled for deprivation using the Under-Privileged Area (UPA) score which explained 43 percent of the variation.

Again some concerns may be raised about the choice of indicators and their coverage as well as the sensitivity to the rankings of health authorities from the transformation of the data.

Health authorities have now been superseded by strategic health authorities and the composite indicator system now in pace is the star rating system for NHS provider organisations. This system is described in a subsequent section.

4.1.4. The World Health Report 2000

The composite index of health system performance produced for 191 countries by the World Health Organisation (WHO) in the *World Health Report* 2000 has been the subject of much debate (World Health Organisation, 2000). The indictor was based on a weighted sum of attainment across 3 broad areas: health, financial fairness and responsiveness. Five dimensions were captured in total as the health and responsiveness indicators considered both the overall level of attainment and their distribution:

- Overall health outcomes
- Inequality in health
- Fairness of financing
- Overall health system responsiveness
- Inequality in health system responsiveness

The index has been widely discussed in the literature (Williams, 2001; Smith, 2002; Appleby & Street, 2001; Nord, 2002) and is described only briefly here. The first dimension average population health was measured in terms of disability-adjusted life expectancy (DALE) that involves the calculation of severity weights for illnesses. The latter used the responses from an international panel of health personnel. The first dimension equity of health was measured primarily in terms of equality in survival for the first 5 years of life. The third dimension fairness in financing was measured by creating an index ranging from 0 to 1, defined as the ratio between total expenditure on health (tax plus out-of-pocket) and total non-food expenditure. The fourth dimension responsiveness was meant to capture how well the health care system responded to basic non-health expectations in the population and was expressed as an index covering seven aspects of responsiveness (dignity and autonomy of patients, freedom of choice of provider, quality of amenities, prompt attention, confidentiality, access to social support networks). Each dimension was scored by around 2000 "key informants" from 35 countries who answered questions about their own country and were then asked to give a score for the aspect as a whole. The seven aspects were then ranked in order of importance by 1000 people and weights assigned based on the rankings. Mean scores on each aspect were multiplied by weights and summed to give an overall responsiveness score. For the other countries that were not asked directly to answer the responsiveness questions, scores were calculated by using the means of variables that had been shown to correlate strongly in the 35 countries. The fifth dimension equity in responsiveness was calculated by asking informants to make judgements about the subgroups they thought were treated with less responsiveness than others. Scores were assigned to sub-groups based on the number of times the country informants mentioned them, multiplied by the share of that group in the population. The products were summed and transformed to give an overall score. Scores for countries other than the 35 questioned were estimated in a similar way to the levels of responsiveness.

Finally, the scores on each of the 5 dimensions were transformed to a 0-100 scale using the formula in the table below and summed using the weights shown in the table, reflecting their relative importance, based on views of about 1000 people from 123 countries, half of whom were WHO staff.

Table 2: Weights and transformations used for the five objectives by WHO (2000)

Objective	Weight	Transformation
H: Overall health outcomes	0.250	(H-20)/(80-20)*100
HI: Inequality in health	0.250	(I-HI)*100
FF: Fairness of financing	0.250	FF*100
R: Overall health system responsiveness	0.125	(R/10)*100
RI: Inequality in health system responsiveness	0.125	(I-RI)*100

Source: World Health Organisation (2000)

A second index was also created using the WHO composite measure of attainment to estimate overall health system efficiency. In this second stage of econometric modelling (using stochastic frontier analysis) an attempt was made to capture relative performance by looking at the difference between what was actually achieved in each country (attainment), compared with the maximum they could be expected to achieve given their resources. The latter was measured by health system expenditure and exogenous influences (human capital captured by years of education). A minimum level was also set and the index of relative performance was calculated by dividing the difference between actual attainment and minimum level by the difference between the maximum and minimum levels.

The debate about the appropriateness of the WHO rankings has been widespread and has touched on most of the potential disadvantages concerning composite indices listed earlier in section 2. Some of these relate to the political context in which WHO operates, the ideological values underpinning the production of the rankings and their purpose (Navarro, 2000; Navarro, 2001; Navarro, 2002; Williams, 2001).

Another set of criticisms relate to the lack of transparency about the methods used, with many commentators noting that it is not easy to find out from the WHO documentation how data were collected and transformed or the characteristics of respondents whose views were taken into account for the measures of responsiveness and the weights (Almeida *et al*, 2001).

The major methodological concerns about the creation of the composite measure have revolved around the nature of the underlying data; the transformation of the data; and the weighting system; and how the weights were elicited (whether respondents' valuations of achievement against objectives truly represent marginal valuations). From an econometric perspective there are also contentious issues around the choice of functional form, the choice of error structure, the choice of covariates and the treatment of exogenous factors. The stochastic frontier method can be highly sensitive to model specification (Jacobs, 2001; Gravelle *et al*, 2004). A huge number of technical and analytical judgements were therefore made in the WHO rankings and whilst many have argued about the potential dysfunctional consequences of poor quality analysis and resultant inappropriate policy responses, the publicity and debate around the WHO exercise put system performance and the methodological issues around composite performance indicators, much higher on the agenda of policy makers and analysts.

4.1.5. The UK star rating system for NHS providers

The NHS Performance Assessment Framework (PAF), published in April 1999, introduced a new broader-based approach to assessing performance in the NHS by encouraging action across six areas (Health improvement; Fair access; Effective delivery of appropriate health care; Efficiency; Patient/carer experience; and Health outcomes of NHS care). The PAF was supported by the publication in June 1999 of the first set of High Level Performance Indicators (HLPIs) and Clinical Indicators (CIs) for both Health Authorities and NHS Trusts respectively (Department of Health, 2000). This was the first full range of indicators for NHS hospital Trusts and gave impetus to the process of publishing information on the performance of NHS organisations in order to provide comparisons and improve performance overall.

In September 2001, the first set of performance star ratings were published by the Department of Health for acute NHS Trusts 2000/01 (Department of Health, 2001). The star ratings are a composite index score given to each NHS organisation which are supposed to provide an overall assessment of performance across a number of indicators. In July 2002, the second set of star ratings were published by the

Department of Health, now covering acute Trusts, specialist Trusts, ambulance Trusts and indicative ratings for mental health Trusts for 2001/02 (Department of Health, 2002a). Primary Care Trusts (PCTs) received a separate publication, describing their performance against a range of suitable indicators, but not a rating. In July 2003, the most recent set of star ratings were published, covering again all types of NHS Trusts and PCTs. In this third round, the Commission for Health Improvement (CHI), the independent regulator of NHS performance, took over responsibility for performance ratings and indicators from the Department of Health (Commission for Health Improvement, 2003).

The methodology for the three years of star ratings has remained relatively constant, although some important changes have been made to the individual indicators covered. The Trust star ratings comprise similar areas of performance to the PAF which, taken together, should give a balanced view of the performance of NHS hospital Trusts. There were broadly four areas of indicators in 2000/01: Clinical effectiveness and outcomes; Efficiency; Patient/carer experience; and Capacity & capability. In the last two years of star ratings the key areas have been: Key government targets; Clinical focus; Patient focus; Capacity & capability; and CHI reviews.

The NHS Performance Ratings system places NHS Trusts in England into one of four categories:

1. Trusts with the highest levels of performance are awarded a performance rating of three stars;
2. Trusts that are performing well overall, but have not quite reached the same consistently high standards, are awarded a performance rating of two stars;
3. Trusts where there is some cause for concern regarding particular areas of performance are awarded a performance rating of one star;
4. Trusts that have shown the poorest levels of performance against the indicators are awarded a performance rating of zero stars meaning that performance must be improved in a number of key areas.

The key government targets are the most significant factors in determining overall performance ratings. The broader range of indicators make up a 'balanced scorecard' to refine the judgement on ratings and are combined in a complex 6-step process to produce the star ratings. CHI reviews of Trust clinical governance arrangements also play an important role in determining star ratings since three star Trusts need to perform well on all key targets as well as the CHI clinical review.

Performance against key targets is assessed in terms of whether the target has been achieved, whether there has been some degree of underachievement or whether the target was significantly underachieved. Trust performance is considered to be of concern if there are:

- a number of targets with some degree of underachievement
- a smaller number of targets with significant levels of underachievement
- some combination of both

The scores are presented in terms of the symbols 3, -, and X representing “achieved”, “under achieved” and “significantly under achieved” respectively.

The methodology broadly entails transforming the underlying key targets and performance indicators from continuous variables into categorical variables into either 3 or 5 categories. The performance indicators in the Patient, Clinical, and Capacity & Capability focus areas are categorised into one of five performance bands, with 5 points awarded for the best performance and 1 for the worst. The thresholds for deciding the cut-offs are not necessarily the same for each variable. A description for each variables threshold is given in the appendix. The default position is described in the following table and simply splits performance into 5 bands by percentile.

Table 3: Percentile method thresholds for scoring non-clinical performance indicators

Performance falls in:	Band given:	Labelled as:
1st to 10th percentile	1	Significantly below average
11th to 30th percentile	2	Below average
31st to 70th percentile	3	Average
71st to 90th percentile	4	Above average
91st to 100th percentile	5	Significantly above average

Source: Department of Health (2002) http://www.doh.gov.uk/performanceratings/2002/method_acute.html

Individual band scores are combined to produce an overall score per area. All indicators are equally weighted within their scorecard area in such a way as to ensure that despite differing numbers of indicators, each scorecard area carries the same weight.

The clinical indicators are published with confidence intervals which means that performance is split into three bands dependent on whether the organisation's confidence interval overlaps the England average for the indicator.

Table 4: Confidence interval method threshold for scoring clinical performance indicators

Overlap with England confidence interval:	Band given:	Labelled as:
No overlap and organisation is worse than England average confidence interval	1	Significantly below average
Overlap, showing organisation is similar to England average confidence interval	3	Average
No overlap and organisation is better than England average confidence interval	5	Significantly above average

Source: Department of Health (2002) http://www.doh.gov.uk/performanceratings/2002/method_acute.html

Exact thresholds for scoring all key targets and performance indicators are given in the appendix.

The CHI clinical review is used to determine poorly performing (zero star) and high performing (three star) NHS organisations. A CHI clinical governance review assesses the trust across seven components of performance:

1. risk management
2. clinical audit
3. research and education
4. patient involvement
5. information management
6. staff involvement
7. education, training and development

Each component is scored from I to IV. After each review, the Trust prepares an action plan to address areas for improvement identified by the CHI report. It is agreed with CHI, and published. For Trusts whose CHI clinical governance reviews are older, CHI, with the support of strategic health authorities, assesses progress against this action plan. Any significant improvements are taken into account in calculating the star rating.

The role of the CHI clinical governance review has evolved over time. In 2001, the assessment for Star Ratings required only that the organisation had not received a critical review. However, at that time, just 16 acute trusts and two mental health trusts had undergone a CHI review. By contrast, when the 2003 Star Ratings appeared, CHI reviews had been published for 160 acute trusts, 28 mental health trusts, 27 ambulance trusts and 28 primary care trusts. The 2003 NHS Star Ratings for acute, specialist and mental health trusts were adjusted in accordance with the 'Finsbury rules' (Commission for Health Improvement, 2003b). In essence, these involve zero-rating any organisation that is evaluated as having achieved only the lowest standard of performance (level 'I') in five or more out of the seven areas of clinical governance assessed, apparently irrespective of the organisation's performance on key targets or the scorecard. Three stars are awarded only to organisations that have achieved key targets, a balanced scorecard, at least three 'III's and no 'I's in the CHI review (Commission for Health Improvement, 2003b).

A complex six-step process is then imposed whereby a sequential set of decisions on achievement on the various key variables determines the ultimate star rating outcome. The most important driving factors for obtaining the highest rating are the key targets and the CHI review which enters both first and again at the very end of the set of decision steps for Trusts to achieve. It thus implicitly is given the highest weighting in the determination of the star rating.

The six-step process is as follows:

Step 1
Does the CHI review show significant weaknesses (calculated by achieving five or more scores of I across the seven components of a CHI review)?
If YES – Rating is zero stars
If NO – Continue to Step 2

Step 2
The rating is then calculated using performance against the key targets. The number of key targets achieved and significantly underachieved is used to determine the level of rating possible. This is detailed in the table below:

Table 5: Matrix of performance rating based on performance against key targets

Key targets significantly underachieved	**3**				Zero stars	Zero stars	Zero stars
	2			One star	Zero stars	Zero stars	Zero stars
	1		Go to step 3	One star	One star	One star	Zero stars
	0	Go to step 4	Go to step 4	Go to step 3	Go to step 3	One star	One star
		9	**8**	**7**	**6**	**5**	**4**
		Key targets achieved					

Source: Department of Health (2002b)

Step 3
This step assesses the Trust's performance on the balanced scorecard for Trusts with moderate underachievement on the key targets. Is the Trust's performance either outside the lowest 20% for all three areas and within the top 50% for one area, or in the top 50% for two of the three areas?
If YES – Rating is two stars
If NO – Rating is one star

Step 4
This step assesses the Trust's performance on the balanced scorecard, for Trusts with little or no underachievement on the key targets. Does the CHI review show some strength and no weaknesses (calculated as one score of III, and no scores of I across the seven components)?
If No – Rating is two stars
If YES – Continue to Step 5

Step 5
Is the Trust's performance on the balanced score card outside the lowest 20% in all three areas and in the top 50% in two out of three areas?
If YES – Rating is three stars
If NO – Continue to Step 6

Step 6
Is there a current CHI review showing significant strengths and no weaknesses?
If YES – Rating is three stars
If NO – Rating is two stars

There are many decisions in the process which may impact on how many Trusts are accorded a particular rating in each category, in particular, the choice of performance indicators, the decisions on the thresholds for transforming the performance indicators into categorical variables, the decision rules which are applied in the matrix and the resultant implicit weighting given to each of the (groups of) indicators. The sensitivity to these choices (using the data for acute Trusts for 2001/02 as an example) is explored in Section 6 of the report.

The star rating outcome has a significant reward and penalty schedule attached to it since Trusts which obtain a three star rating for a consecutive number of years may apply for Foundation Status which will give them significant financial and managerial decision-making freedoms and autonomy from central involvement. The incentives for ensuring a good outcome on the star rating are therefore very strong.

The ratings are intended to be 'not primarily a commentary on the quality of clinical care', but rather to assess the 'overall patient experience' (Department of Health, 2001). However, both the construction and the impact of Star Ratings have been questioned (Kmietowicz, 2003; Cutler, 2002; Snelling, 2003; Miller, 2002). Many of the concerns that have been raised on their construction are considered in section 6 using the data from the star ratings to explore the relevant methodological issues.

4.2. Other public sectors in the UK

4.2.1. Local government performance assessment

Besides the health sector, a number of other public sectors in the UK use composite indicators in their performance rating systems. For example, the Audit Commission undertakes a Comprehensive Performance Assessment (CPA) of local government that covers a number of core services and draws on approximately 100 performance indicators (DTLR, 2001). CPA has been published in December 2002 and December 2003. This is an extremely complex composite indicator, since it is built up from a number of underlying composite indices measuring the key services.

The key services which are assessed are the following seven domains:

1. Benefits
2. Education
3. Environment
    a. *Transport*
    b. *Waste management*
    c. *Planning*
4. Housing
    a. *Housing management*
    b. *Community Housing*
5. Libraries & Leisure
6. Social services
    a. *Children's Services*
    b. *Adults' Services*
7. Use of resources

The CPA is performed by the Audit Commission who also cover the performance assessment for the domains environment, housing, libraries and leisure and use of resources (Audit Commission, 2003a; Audit Commission, 2003b). Star Ratings for education, social services and benefits are produced as part of the CPA process and contribute to the overall CPA assessment. Education assessments are made by the Office for Standards in Education (Ofsted). Social care assessments are made by the Social Services Inspectorate (SSI) and benefits assessments are made by the Benefits Fraud Inspectorate (BFI).

This section describes how the seven domains (assessed by the four different inspectorates) are combined into the overall CPA scorecard for each council. Subsequent sections describe the construction of the underlying composites for the seven key areas. These all have very large similarities to the Star Ratings in the NHS and are therefore described in some detail.

The CPA is an assessment of each authority’s performance across the full range of its services, its financial integrity, capacity to drive improvement, its underlying systems and processes and the way in which it relates to its community and partners. The scorecard summarises the performance of every assessed council. The overall CPA judgement draws information from a range of Government inspectors reports, Best Value Performance Indicators (BVPIs), audits, and assessments of service plans.

Key services then receive a score ranging from 1 (the lowest) to 4 (the highest). The Audit Commission also assesses how the council is run and how the council is going about delivering its priorities, again using a score of 1 to 4. These two scores are combined to reach the overall score of excellent, good, fair, weak or poor, which is known as ‘the scorecard’.

The following table shows the number of Best Value Performance Indicators used in the CPA.

Table 6: The number of Best Value Performance Indicators (BVPIs) used in the Comprehensive Performance Assessment (CPA)

Corporate Health	16
Education	18
Social Services	14
Housing	9
Housing Benefit and Council Tax	6
Waste	8
Transport	9
Planning	5
Environmental Health and Trading Standards	1
Culture Services / Libraries and Museums	4
Community Safety	7
Community Legal Service	1
Fire	11

Source: (ODPM, 2003)

Because there are differences in the size of councils and the types of services they deliver, two streams of assessment have emerged, namely Single Tier and County Councils and District Councils. Although these differ in the details, the overall approaches are similar (Audit Commission, 2003c).

The scores from all the services are combined according to the following process (for Single Tier and County Councils respectively) as set out in the following table.

Table 7: Determination of scorecard for Comprehensive Performance Assessment (CPA)

1. Current scores on services range from 1 (poor) to 4 (excellent)		
2. Weighting between services on the seven domains is:		
	Weight	
Education	4	
Social services (average of children's and adults)	4	
Environment	2	
Housing	2	
Libraries and leisure	1	
Benefits	1	
Use of resources	1	
3. Maximum and minimum scores are as follows (counties in brackets):		
	Minimum	*Maximum*
Education	4	16
Social services	4	16
Environment	2	8
Housing	2(0)	8(0)
Libraries and leisure	1(0)	4(0)
Benefits	1	4
Use of resources	1	4
Total	15(12)	60(48)
4. Combine core service scores to reach overall judgement:		
	Single tier	*Counties*
1 = lowest	Less than 30	Less than 24
2	30-37	24-29
3	38-45	30-36
4 = highest	More than 45	More than 36
5. Combine score of how council is run from 1 (poor) to 4 (excellent) – weighted average of less than 2 gives overall score of 1 and weighted average of more than 3 gives overall score of 4		
6. Application of rules (to overall category):		
Rule 1: Must score at least 3 (2 stars) on education, social services combined star rating, and financial standing to achieve a category of excellent overall. *Rule 2*: Must score at least 2 (1 star) on education, social services combined star rating, and financial standing to achieve a category of fair or above. *Rule 3*: Must score at least 2 (1 star) on all core services to achieve a category of excellent overall.		

Note: Scores are converted as follows:
Education: 0 star = 1, 1 star = 2, 2 stars = 3, 3 stars = 4
Social services (average score for children and adults): not serving people well = 1, serving some people well = 2, serving most people well = 3, serving people well = 4

Education and social services (which for instance each have a star rating system underpinning them) receive a higher weighting in the overall composite CPA score. In the final step, a set of decision rules are applied (similar to the Trust star ratings in healthcare) which again may impact on the final CPA score given to an authority.

There are significant rewards for high performing councils in the form of:

1. Less ring-fencing
2. Fewer and 'lighter touch' inspections
3. Fewer planning requirements
4. Freedom to use income from fines

The reaction of local government councils to CPA has been mixed with strong support for the self-assessment and peer assessment aspects as well as the financial freedoms, but concern over whether the overall judgements fairly represent performance.

4.2.1.1 Housing, Environment, Use of Resources and Libraries and Leisure

In the CPA, Best Value Performance Indicators (BVPIs) are scored for the service blocks of Housing, Environment, Use of Resources and Libraries & Leisure. The general approach is to award a percentile ranking to each Performance Indicator (PI) for an authority based on an all-England comparator group.

This is a two-step process, as shown in the following table below, based on ten fictitious authorities.

Step 1
The percentile is calculated relative to the all-England comparator group.

Step 2
The raw PI values are turned into negative figures and the percentile is calculated. These two percentiles are then averaged.

Table 8: The percentile approach for scoring performance indicators

	Raw PI	Step 1	Negative Raw PI	Step 2	Average
Authority 1	29.4	0	-29.4	0.000000	0
Authority 2	41.0	0.222222	-41.0	0.222223	0.22
Authority 3	42.0	0.333333	-42.0	0.444445	0.39
Authority 4	56.3	0.777777	-56.3	0.777778	0.78
Authority 5	38.0	0.111111	-38.0	0.111112	0.11
Authority 6	63.2	1	-63.2	1.000000	1.00
Authority 7	42.0	0.333333	-42.0	0.444445	0.39
Authority 8	45.3	0.555555	-45.3	0.555556	0.56
Authority 9	63.0	0.888888	-63.0	0.888889	0.89
Authority 10	48.8	0.666666	-48.8	0.666667	0.67

Source: Audit Commission (2003a)

In this example, a high value on the raw PI is desirable. The highest performer (Authority 6) gets a percentile of 1 (equivalent to 100th percentile), and the worst, Authority 1, gets a percentile of 0 (i.e. zero percentile).

Within each domain or service block, there is a basket of PIs and the percentile scores for each PI in the block are averaged to give an average percentile for the service block, as shown in the example in the following table for Use of Resources.

Table 9: Averaging the percentiles for the service block: Use of Resources - Financial Administration

	BVPI 8	BVPI 9	BVPI 10
Percentiles for Authority 10	0.49	0.26	0.62
Average percentile for this service block	**0.46**		

Source: Audit Commission (2003a)

This average percentile is scored 1 - 4 (with 4 as the highest score) using the thresholds set out in the following table. These thresholds reflect the fact that it is more difficult to achieve a good performance on every PI in a basket when there are a large number of PIs. Therefore, where an authority has an average percentile equivalent to having three-quarters of its PIs in the top quartile for any service block, and none in the lowest quartile, it will score a 4 and vice versa to score a 1.

Table 10: Converting average percentiles to scores

Average percentile of:	Scores:
above 0.625	4
above 0.5, below or equal to 0.625	3
above 0.375, below or equal to 0.5	2
equal to or below 0.375	1

Source: Audit Commission (2003a)

However, sometimes there are fewer than four PIs within a block in which case thresholds are adjusted as follows:

Table 11: Adjusted thresholds when there are fewer than 4 PIs

Scoring a 1 when there are fewer than 4 PIs		Scoring a 4 when there are fewer than 4 PIs	
Number of PIs	Adjusted threshold	Number of PIs	Adjusted threshold
1	0.25	1	0.75
2	0.31	2	0.69
3	0.33	3	0.67

Source: Audit Commission (2003a)

To take year on year changes into account, percentiles are 'pegged' at 2000/01 values. Thus if in 2000/01 a performance of 52% on a PI meant that an authority was awarded the 80th percentile, a performance of 52% in 2001/02 would attract the same, 80th percentile. If the percentiles were not pegged, a performance of 52% in 2001/02 may have attracted a percentile of only 75th, say, if many authorities had improved their performance. Thus, pegging provides an opportunity for all authorities to improve their PI score year on year.

Authorities' performance is scored against an all-England comparator group in most cases, except for a few exceptions:

1. *Standards*. Where statutory standards of performance exist, the percentiles are adjusted so that the top (100th) percentile is achieved if the standard is met.
2. *Lack of variation*. When there is little variation in the PIs, percentiles are not used, since very small differences in performance would attract quite large differences in percentiles. Instead, quartiles are used. However, where small differences in performance represent significant differences in impact percentiles are still used.
3. *Dichotomous PIs*. Scores of 0.25 (no) or 0.75 (yes) are used.
4. *Scoring against contextual data*. Some PIs are considered alongside the local circumstances of the authority rather than just compared nationally.

Some of the performance indicators are adjusted for social deprivation. The measure of deprivation used was the Average Ward Score of the Index of Multiple Deprivation (IMD) (DETR, 2000). Linear regression analysis was used to examine the relationship between social deprivation and various performance indicators and adjustment were made where a relationship was found to be statistically significant, and considered to be a causal one. User satisfaction PIs were thus adjusted for social deprivation, using the following formula:

$$\textit{Expected satisfaction level} = \textit{constant} + \textit{coefficient} \times \textit{IMD} \qquad (1)$$

where IMD is the Index of Multiple Deprivation of the authority. The expected performance figure is subtracted from the actual performance to give a residual, which is then ranked by percentiles and used in the PI analysis. Using this equation, the coefficient in every case turns out to be negative, reflecting the fact that satisfaction tends to decline as deprivation increases.

4.2.1.2 Benefits

The Benefit Fraud Inspectorate (BFI), part of the Department for Work and Pensions, undertakes an assessment of the Housing Benefit and Council Tax Benefit service provided by each district council. This is in many cases not a full inspection but an evaluation based on self-assessment. BFI will then use this plus other performance information including BVPIs to produce an overall assessment and report. The BFI report will give an overall rating for both current performance and for capacity to improve. The ratings are based on a five-point scale (poor, fair, fair to good, good, and excellent).

The Audit Commission uses a 4 point scale on the CPA scorecard, hence the BFI's assessment is translated to the Audit Commission's service scorecard using the thresholds shown in the following table.

Table 12: Benefit Fraud Inspectorate assessments within the CPA

BFI assessment	Audit Commission's scorecard
Excellent, Good, Fair towards good (80% or above)	4
Fair (60-79%)	3
Fair (40-59%)	2
Poor (0-39%)	1

Source: Benefit Fraud Inspectorate (2003)

4.2.1.3 Social services

The Social Services Inspectorate (SSI) (soon to be superseded by the Commission for Social Care Inspection) produces Personal Social Services (PSS) Star Ratings that assess the current performance and prospects for improvement of social services in the areas of services for children and adults (Department of Health, 2002c; Department of Health, 2003; Commission for Social Care Inspection, 2003).

In May 2002, the Department of Health SSI published the first set of social services PSS star ratings. These covered all councils with social services responsibilities in England using all the evidence available at that time. A second updated set of star ratings was published in November 2002, including more up to date performance indicators and inspections, followed by the third set of star ratings published in November 2003.

These performance ratings have been formulated from evidence from published Performance Indicators, inspection, Social Services Inspectorate (SSI) / Audit Commission Joint Reviews, self-assessment, and reviews of plans and in-year performance information from both the SSI and the external auditors for each council.

The social services star rating feeds directly into the local government CPA. A council must receive a good star rating for their social services in order to receive the highest comprehensive performance assessment rating.

The performance indicators are selected against the following four criteria:

- *Importance* – clearly relating to government priorities;
- *Ease of interpretation* – not requiring further contextual data to understand, with clear criteria to identify good and bad performance;
- *Reliable data* – the data provided by councils are believed to be reliable and definitions of indicators sufficiently capture good practice;
- *Attributable to social services* – the level of the indicator is largely due to the performance of social services, rather than other factors or agencies.

Domains include meeting national priorities, cost and efficiency, effectiveness of service delivery, quality, fair access to services, and prospects for improvement.

Judgements for children and adults services are given. In both cases, a judgement for both current performance and prospects for improvement is given. This results in a total of four judgements underpinning the overall rating. Once the judgements have been reached, a set of decision rules is used to combine them with the weightings to produce a final star rating.

The principles underlying the decision rules are as follows:

- current performance is weighted more heavily than prospects for improvement;
- adult services and children's services are given equal weight;
- a "failure" in either adult services or children's services will result in zero stars, no matter how good the other services are.

A subset of performance indicators are defined as the Key Performance Indicators and are each given a threshold value determining the maximum judgment that can be given to the indicator. For these, a council could not be judged to be performing well if it failed to reach a specified band of performance. There are minimum standards for both children and adult performance indicators, and a council will have to meet all the minimum standards in order to receive one of the higher judgments.

The following table shows how the star ratings are presented.

Table 13: Star ratings for social services performance

	Performance rating	Children's services		Adults' services	
		Current performance - Serving people well?	Improvement prospects?	Current performance - Serving people well?	Improvement prospects?
Council 1	-	No	Poor	Most	Promising
Council 2	=	Some	Uncertain	Some	Promising
Council 3	= =	Most	Promising	Yes	Uncertain
Council 4	= = =	Most	Excellent	Yes	Promising

Source: Department of Health (2002c)

Social services are provided or arranged by local councils, but are often planned and delivered in partnership with the NHS and other council services. The social services star rating is therefore designed to be compatible with performance information for both the NHS and other local government services.

4.2.1.4 Education

The Office for Standards in Education (OFSTED) is a non-ministerial government department independent of the Department for Education and Skills (DfES). OFSTED produces an ‘education profile’, or scorecard, with two separate Star Ratings for each local education authority (LEA) in the areas of current performance and improvement (as well as capacity to make further improvements), similar to social services. Five domains are assessed: school improvement, special educational needs, social inclusion, lifelong learning and strategic management of education (OFSTED and DfES, 2002). The league tables produced by the Department for Education and Skills for primary and secondary schools and colleges in the UK contribute to the CPA through the ‘education profile’.

All schools are inspected at least once within six years. Inspectors make judgements on a seven-point scale as follows: Excellent 1, very good 2, good 3, satisfactory 4, unsatisfactory 5, poor 6, and very poor 7. The Evaluation Schedule applies to the inspection of all nursery, primary, secondary and special schools, pupil referral units and any other publicly funded provision (Office for Standards in Education, 2003). The Evaluation Schedule covers the following key areas:

1. Effectiveness of the school
2. Standards achieved by pupils
3. Quality of education provided by the school
4. Leadership and management of the school

Most elements of the education profile are made up by combining performance indicators and inspection judgements. The education profile has the potential for fifteen elements or assessments to be made (across the three areas - current performance, improvement, capacity; and five domains - school improvement, special educational needs, social inclusion, lifelong learning, strategic management). Each of the elements across the five domains is then aggregated to give an overall assessment score for each of the three areas.

There are 45 indicators used to feed into the elements of the profile. Of these, 24 are performance indicators and 21 are LEA inspection judgements. The performance indicators show how well an LEA is doing compared to all LEAs. The inspection judgements show how well the LEA is doing compared to the standards set in OFSTED’s criteria for inspection judgements.

No adjustment is made to the profile for social deprivation. This is for two main reasons. Firstly, OFSTED argue that national funding is designed to recognise the challenges faced by an LEA. Secondly, nearly half (21 out of 45) of the indicators used are based on inspection judgements, which are made taking the context of an LEA into account.

Each indicator is converted to a categorical score on a five-point scale with 1 being the highest and 5 being the lowest score. For the performance indicators, all LEAs are ranked and the score is then determined by allocating the top 10% a 1, the next 20% a 2, the next 40% a 3, the next 20% a 4 and the remaining 10% a 5. For inspection judgements, which are made in the first instance on a seven-point scale, the inspection grades are converted to a five-point scale.

The scores within each of the fifteen elements are then added together and divided by the number of scores to give an overall score for the element. This is the score shown on the education score card for the element.

A few weights have been introduced into the creation of the profile with respect to the construction of each element and the weights attached to certain indicators and inspection judgements. Otherwise, there are no other weightings in the profile. Each element has an equal effect on the overall score for each area.

The Audit Commission model uses scores on a four-point scale to feed into the CPA for a council. To this end, the average score, and the ranked position, of an LEA are used to determine its category on the four-point scale.

The ranked list for current performance is allocated star ratings on the basis of the inspection grades and performance quotas for each category. The improvement categories are derived differently and the rules used to allocate star ratings for improvement are shown in the following table.

Table 14: Operation of performance and improvement rules

Performance stars	Improvement score	Indicated improvement	Improvement stars	Capacity score	Indicated capacity	Improvement stars
= = =	1.0 - 2.9	Proven	= = =	-	-	-
= = =	3.0 - 5.0	Not proven	-	1.0 - 2.9	Secure	= =
= = =	3.0 - 5.0	Not proven	-	3.0 - 5.0	Not secure	=
= =	1.0 - 2.9	Proven	= = =	-	-	-
= =	3.0 - 5.0	Not proven	-	1.0 - 2.9	Secure	= =
= =	3.0 - 5.0	Not proven	-	3.0 - 5.0	Not secure	=
=	1.0 - 2.9	Proven	-	1.0 - 2.9	Secure	= =
=	1.0 - 2.9	Proven	-	3.0 - 5.0	Not secure	=
=	3.0 - 5.0	Not proven	-	1.0 - 2.9	Secure	=
=	3.0 - 5.0	Not proven	-	3.0 - 5.0	Not secure	-
-	1.0 - 2.9	Proven	-	1.0 - 2.9	Secure	= =
-	1.0 - 2.9	Proven	-	3.0 - 5.0	-	=
-	3.0 - 5.0	Not proven	-	1.0 - 2.9	Secure	=
-	3.0 - 5.0	Not proven	-	3.0 - 5.0	Not secure	-

Source: Office for Standards in Education (2002a)

The following table shows an example of a LEA education profile, for the city of York.

Table 15: Example of Local Education Authority CPA education scorecard (for the City of York)

Aspect	Current performance	Indications of improvement	Capacity to make further improvement
School Improvement	2.1	2.4	1.6
SEN	2.5	1.0	4.0
Social Inclusion	1.6	2.8	2.3
Life Long Learning	2.7	2.0	2.0
Strategic Management	1.3	-	2.5
Average Score	2.0	2.3	2.2
Category	= = =	= = =	

Note: The Average Score thresholds for the Performance star ratings are as follows:
3 star is obtained if the Average Score is less than or equal to 2.38
2 star is obtained if the Average Score is less than or equal to 3.34 but more than 2.38
1 star is obtained if the Average Score is less than or equal to 3.75 but more than 3.35
- star is obtained if the Average Score is greater than 3.75

Source: Office for Standards in Education (2002b)

The whole CPA process is therefore hugely complex and built on a structure of underlying composite indices, each with a huge number of methodological choices underpinning them. Some of the potential pitfalls with the process will be discussed in subsequent sections, but these include the application of decision rules to construct the star ratings, the widespread use of categorical variables with potentially arbitrary thresholds, and the application of percentile thresholds, the opaque use of weights, and the inconsistency in dealing with exogenous factors. One of the main differences though with the star rating system applied to health care is the reward schedule attached to the CPA. Authorities have to deal with more inspections if they perform poorly, whereas in NHS hospital trusts, the management team can effectively be replaced.

4.2.2. Performance assessment in universities

Performance assessment in universities takes the form of two independent assessments for teaching and research at higher education institutions. These do not take the form of composite indices as described above, but do share some common features in that they consist of an aggregation of underlying performance information which culminates in a judgement on a rating scale. University departments are quality rated by the Quality Assurance Agency for Higher Education for teaching (subject reviews, in which grades are awarded in a specific subject) and research (Research Assessment Exercise, RAE) (HEFCE, 2003).

4.2.2.1 Teaching

In 1997, the Quality Assurance Agency for Higher Education (QAA) was established to provide an integrated quality assurance service for UK higher education (The Quality Assurance Agency for Higher Education, 2003). The QAA is an independent body funded by subscriptions from universities and colleges of higher education, and through contracts with the main higher education funding bodies.

Each higher education institution is responsible for ensuring that appropriate standards are being achieved and a good quality education is being offered. The QAA safeguards public interest in standards of higher education qualifications, by reviewing standards and quality, using a peer review process where teams of academics conduct audits and reviews.

The review period extends over a period of about six weeks. During the visit, the review team gathers evidence to form judgements on the standards and quality of the provision of teaching. This is achieved through scrutiny of documentary evidence, meetings with relevant staff and, sometimes, direct observation of teaching. The review team meets current students during the visit, and they may also meet former students and their employers from relevant industries or professions.

For each academic review, the team expresses 'confidence', 'limited confidence', or 'no confidence' in:

- academic standards (learning outcomes; the curriculum; student assessment; student achievement).

The team also makes judgements of 'commendable', 'approved' or 'failing' for:

- the quality of learning opportunities (teaching and learning; student progression; learning resources).

The published report sets out the review team's judgements. It also states the degree of confidence the team has in the institution's ability to maintain and enhance quality and standards in the subject under review.

4.2.2.2 Research

Employing a seven-point scale, the main purpose of the RAE is to enable the higher education funding bodies to distribute public funds for research selectively on the basis of quality. Institutions conducting the best research receive a larger proportion of the available grant (around £5 billion of research funds) so that the infrastructure for the top level of research in the UK is protected and developed. Very strong financial incentives are therefore associated with RAE scores.

To judge the quality of research, HEFCE conduct a Research Assessment Exercise (RAE) every four to five years. The most recent exercise took place in 2001. Institutions were able to submit research outputs in up to 69 subject areas (units of assessment), which are discipline-based. Each submission consists of information about the academic unit being assessed, with up to four publications and other research outputs for each member of research-active staff. All forms of research output (books, papers, journals, recordings, performances) are treated equally. Similarly, all research (whether applied, basic or strategic) is treated equally. The range of outputs might include publications, products or even artistic performances, assessed through peer review by panels of experts in the subject. For each unit of assessment there is a panel of between nine and 18 experts, mostly from the academic community but with some industrial or commercial members as well. There are 60 assessment panels; usually there is one panel for each unit of assessment but a few units of assessment have joint panels. Panel members are nominated by a wide range of organisations, and then selected by the funding bodies, on the advice of the panel chair.

The panels each draw up a statement describing their assessment criteria which is published in advance of submissions being made. This statement shows which aspects of the submission the panel regards as most important, and areas that it wants institutions to comment on in their submissions. Panels do not visit institutions as part of their work (HERO, 2001).

Every higher education institution in the UK may make a submission to as many of the units of assessment as they choose. Submissions are made in a standard format, which includes qualitative and quantitative information. Each submission is given a quality rating, judged against standards of national and international excellence, on a seven-point scale, from 1 at the bottom through 2, 3b, 3a, 4 and 5, to 5* (five star) at the top. Outcomes are published and so provide public information on the quality of research in universities and colleges throughout the UK. However, unlike the school league tables, there is no equivalent official ‘performance table’ for universities.

The policy of selective funding for research through the process of research assessment has contributed to improvements in the quality of research in the UK. A recent study commissioned by HEFCE into the role of selectivity concluded that (HEFCE, 2003):

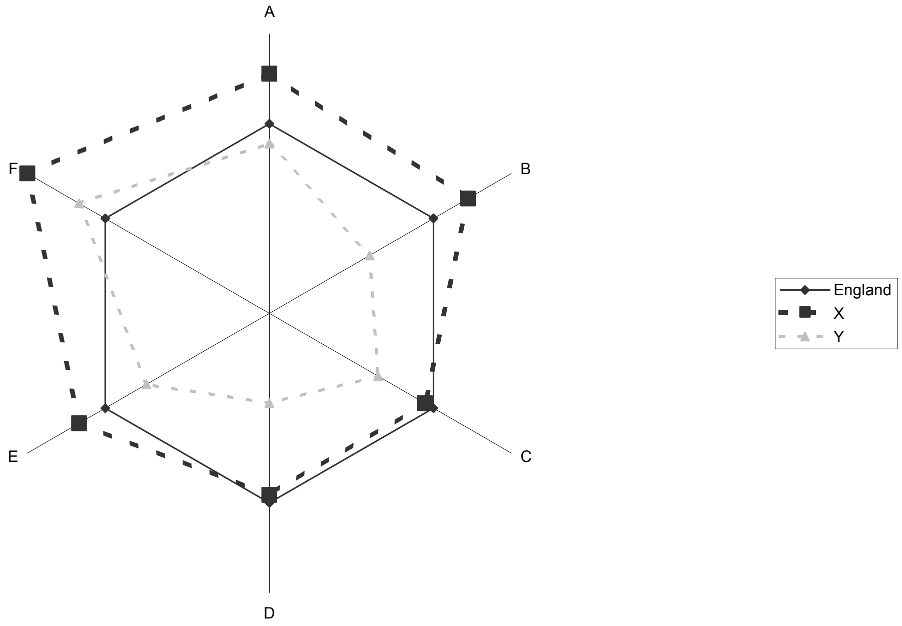
- in the period since the first RAE in 1986 the effectiveness and productivity of the UK research base has increased substantially. UK researchers are among the most productive, and the number of times their work is read and used by other academics per million pounds spent is the highest in the world;
- research activity in the UK has increased at a faster rate than funding, indicating an increase in efficiency;
- the introduction of a national system for the assessment of research quality has been effective in improving the management of the research environment.

4.2.3. Policing performance assessment

Policing is one sector where the construction of a composite indicator has effectively been avoided. Police performance is monitored as part of the Policing Performance Assessment Framework (PPFA). This framework for assessment is still in the early stages of development. The five domains of policing activity which are assessed include reducing crime, investigating crime, promoting public safety, citizen focus, and resource use. The key performance indicators referred to in the National Policing Plan 2003-2006 are grouped together into these five domains, each containing like indicators. The indicators in each domain are aggregated together, most often applying an equal weighting to the underlying indicators, but not always. Under investigating crime, because of the scaling of the underlying indicators, and the way they are combined, 'Class A drug offenders brought to justice' effectively receives a five-fold higher weighting. A spider diagram or 'radar' plot can be produced for each unit under assessment which summarises performance pictorially on each of the domains, rather than attempting to produce a composite indicator (Home Office, 2002). The spider diagram avoids the need to arbitrarily apply a weighting structure to each of the domains, although weights are still explicitly or implicitly applied to the underlying performance indicators in each domain. The down-side to applying this type of visual approach is that only a limited number of indicators can realistically be shown in a single plot.

The diagram in Figure 1 is comprises of six scales, one for each performance indicator (A to F) which all need to measure performance in the same direction i.e. 'more is better'. The England hexagon shows average performance across all units on each of these dimensions or scales. Organisation X therefore underperforms on all indicators except one and in general those organisations closer to the centre will exhibit poorer performance on the indicators. Organisation Y is performing better than the England average across most of the performance indicators and such organisations will tend to approach the extremes of each scale. The figure essentially shows a snapshot of performance, but in principal performance year on year trends could be incorporated into this type of analysis.

Figure 1: Example of radar plot to show comparison across multiple indices



The approach typically used in policing is to compare forces to a sub-set of other similar comparable units rather than (say) the England average, to make comparisons of like units more meaningful (sub-group analysis).

Although this form of performance assessment in policing does not take the form of an overall composite index and produces an alternative way of presenting the different dimensions of performance, there are still similar potential pitfalls as in the construction of composites, since each dimension still requires careful judgement over the choice of indicators, their transformations, their weighting, and how to combine them in a single domain.

5. The steps to developing a composite indicator

In this section the methodological issues involved in constructing a composite indicator are explored, illustrating where appropriate with examples from the literature.

There are a number of steps involved in constructing a composite indicator, each of which requires important judgements to be made by the analyst. At each step, these judgements can potentially have a substantial impact on the ultimate outcome of the composite and therefore require careful consideration:

- Choosing the entities to be assessed;
- Choosing the organizational objectives to be encompassed in the composite;
- Choosing the indicators to be included in the composite;
- Transforming measured performance on individual components;
- Combining the components, using some sort of addition or other decision rule;

- Adjusting for environmental or other uncontrollable influences on performance;
- Adjusting for variations in expenditure (if a measure of efficiency is required);
- Conducting a sensitivity analysis to test the robustness of the outcome of the composite to the various methodological choices.

These steps are examined in more detail below.

5.1. Choosing the units to be assessed

This is not a trivial matter and hinges on the decision about what the boundaries are of the units to be assessed and what aspects of performance these units will be held responsible for. In health care, these boundaries are often blurred, not least by the fact that there are increasingly performance incentives to ensure a seamless service provision across different organisational boundaries, such as between primary and secondary care and between secondary and residential or long-term care and between social services and health care, and so on.

The choice of the organisational units will also impact on the choice of indicators which are available with which to assess their performance. The degree of homogeneity across units will also have an effect on the choice of indicators and how uniform these can be. If units are very heterogeneous, there may be a demand from the units under assessment to include a broader range of indicators on which each of the different units can be given a fairer rating.

Such problems are more likely to arise where, for instance, composites are created to compare across rather than within countries. The WHO health system performance indicator included all countries in the world but the purpose and usefulness of comparing health systems in such disparate countries as the USA and Malawi is questionable. Again the choice of indicators may be more relevant to the circumstances of some countries than others and it is likely that the meaning of some of the dimensions chosen (such as autonomy and dignity) varies widely between countries.

There are other examples of composite indictors created at an international level, such as the Environmental Sustainability Index created by the World Economic Forum (2002) which attempts to measure for 142 countries, the overall progress towards environmental sustainability using 20 sub-indicators and 68 variables. Similarly, the United Nations Human Development Index attempts to measure human development along 3 dimensions (life expectancy at birth; literacy; and GDP per capita) for all countries in the world (United Nations, 2003). The issue of potential heterogeneity between units of assessment is less problematic where coverage is limited to “similar” organisations or where the dimensions measured are relatively straightforward to measure and not subject to a great deal of variation in definition. Thus many composite indictors have been restricted to sub-sets of countries such as those in Europe or the OECD, for instance the Composite of Leading Indicators (OECD) (Freudenberg, 2003); European labour market performance (Storrie and Bjurek, 2000); and many use indicators that are tightly defined and applicable across smaller groups of countries (such as unemployment and inflation). Similarly, the heterogeneity of countries in the WHO exercise, has lead to the examination of latent class models and natural groupings of units (for example Sub-Saharan Africa) with similar characteristics within which to compare health system performance (Greene, 2003). This is similar to the type of sub-group analysis done in policing.

5.2. Choosing the organisational objectives to be encompassed

Prior to making a choice of individual indicators to include in the composite, a decision must be made about the overall objectives against which the organisations are to be assessed. This is probably mainly a political task. In England, the Public Service Agreements are the main instruments used to signal the strategic objectives for agents in the public service.

Where the composite indicator seeks to measure something very specific such as dental health (Marcenes and Sheiham, 1993) or "innovation capacity" (Porter and Stern, 1999), the setting of objectives may not be as problematic as more sweeping all-embracing indicators.

The WHO index attempts to measure performance of the entire health system across 191 countries and it is clear that the choice of appropriate indicators to make up the composite is fraught with problems. Health care systems differ greatly in their objectives (Navarro, 2000) and operate in very different political and economic contexts (Nord, 2002).

Much of the criticism of the WHO measures of health system performance arises from disagreement about the appropriateness of the dimensions against which performance should be assessed which reflect in part the disagreement about the objectives of different health care systems.

5.3. Choosing the indicators to be included

This step is probably one of the most fundamentally important where judgement is required. In practice, composites are often either opportunistic and incomplete (measuring aspects of performance that are captured in existing data), or are based on highly questionable sources of data. Either weakness can cause serious damage to the credibility of the composite (Smith, 2002). The choice of indicators is most often constrained by data availability and may give an unbalanced picture of health services.

A different set of indicators will produce a different composite indicator and hence a different set of rankings, although it is not known how different. Hence the choice of indicators is absolutely imperative. The indicators which are included in the composite, the weight which is attached to them, and the reward schedule which is attached to the outcome, will have an effect on the effort which will be expended on trying to achieve on the included indicators, at the potential expense of the excluded indicators. These excluded indicators may be as (or more) important than the included indicators but simply more difficult to measure.

There are also issues about whether the organisational units which are being measured are able to understand and replicate the construction of the included indicators. For instance if some indicators are constructed as comparators to other relevant units or, say, the national average, the individual units will not have access to much of this data. An inability to understand and duplicate the construction of the indicators, may make it difficult for the units to take corrective action where needed.

Aside from data availability, a number of other issues arise in the choice of performance indicators, namely, the types of indicators to be included, data reliability, and the collinearity between the indicators.

5.3.1. Process or outcome measures

There has been much debate about the different types of performance indicators in health care, in particular between the choice of process versus outcome measures. Outcome measures assess whether medical care has achieved certain results, while process measures assess whether medicine has been properly practised. One can seldom be confident that outcome measures such as current health status is an indicator of current health system performance. Current experience of the health system will not have a contemporaneous effect on health status (Donabedian, 1966). Thus it is argued that measures of health system process rather than health status outcome may be preferable.

Process measures relate to what is done for the patient and assess degrees of conformity to various 'accepted standards' of care. Process measures might include utilisation rates for laboratory procedures, referral rates, autopsy rates, technical competence in diagnostic and therapeutic procedures, continuity of care and average length of stay by diagnosis. In some circumstances such as the management of chronic diseases, process measures will be far more appropriate than outcome measures (Crombie and Davies, 1998). Since process measures tend to be very specific, the desire to be comprehensive in the construction of a composite, may lead to a lengthy set of indicators being included.

Outcome measures look at the effects of hospital or other forms of health care on patient health status. These include general mortality, morbidity and disability rates, illness and health status indices, case fatality rates, readmission rates, complication and infection rates. They can be grouped in terms of death, disease, disability, discomfort and dissatisfaction. Outcome measures are less vulnerable to reporting errors or misrepresentation by health care professionals.

While the outcome approach may be conceptually appealing it is empirically problematic and can be rejected on other grounds. These include considerations such as the fact that health care systems do not necessarily produce health. Furthermore, if this notion was followed and no health status change was effected one would have to assume zero health system output. On the other hand, it is worth considering that managers, patients and carers will be most concerned with effecting a change in health status (Butler, 1995). The patient does not desire "two office visits, five days of hospital care, three X-rays, and sixteen tablets of antibiotics, but rather the expectation that his level of health will be improved" (Berki, 1972 pg. 32).

One of the main problems with outcome measures is distinguishing between those changes which are the result of treatment and those which result from other factors, such as the natural progression of the disease, age, sex, socio-economic status of the patient, and behaviour and compliance with a treatment regime. This relates back to the extent to which health status can be attributed solely to the health care system. If the definition of the health care system is narrowly confined to the role of certain agencies and does not include broader concepts such as lifestyle and diet, then the influence of these factors on the health care system (and the performance assessment of these agencies) could be profound.

If an attempt is made to use more sophisticated measures of health outcome, there is a lack of consensus about the most appropriate indicators to use. For example, the WHO index used Disability Adjusted Life Expectancy (DALE) in order to capture both mortality and morbidity, giving rise to much debate about the appropriateness of this particular measure of health outcomes (Williams, 2001; Almeida *et al*, 2001), as well as the specific version adopted by the WHO (Nord, 2002). Others highlight the difficulties of introducing a complex measure requiring substantial data manipulation and "heroic assumptions" into an already complicated process of creating a composite indicator of performance (McKee, 2001). There may therefore be arguments for using simpler process measures.

While there are no hard and fast rules as to whether to adopt one or the other measure, there is often a spectrum between immediate process and eventual clinical outcome, with a variety of intermediate measures of outcome (Davies *et al*, 2002).

5.3.2. Types of indicator variables

Indicator variables may be set up in a number of different ways depending on what they are measuring. Some common types of performance indicators include threshold variables or dichotomous indicators, continuous variables and indicators with confidence intervals, and change indicators.

Threshold variables are indicators which measure service standards and have a fixed threshold for their attainment. Performance might be measured against government targets which specify, for instance that no patients should wait longer than 18 months for inpatient treatment. These types of variables typically display little variation, especially if severe penalties are attached to a breach of the government target. They are not therefore always well suited to picking up differential performance, since they measure attainment against a standard rather than against other organisations and really amount to a dichotomous variable.

Within the NHS star rating system, the Key Targets often take this format (rather than the balanced scorecard types of variables which display more variation). There are strong incentives for managers to invest greater effort in first attaining these Key Targets, before continuing to achieve on the balanced scorecard. Managers will therefore often 'overshoot' in their efforts on these threshold variables, since if they do fail on these indicators, it makes no difference by how much and the thresholds could potentially be sensitive to small-number variation.

However, lack of variation may not be a reason to exclude these types of variables from the composite, since the actual performance they are measuring may represent important policy goals (for instance long waits or cleanliness). If the threshold is all that matters, then lack of variation is not a problem in itself, if all units achieve the standard. However, if detecting greater performance variation is important, then there may be an argument to try and measure these aspects of performance in a different way, or simply change the threshold values to pick up more variability. For example, the measurement of hospital cleanliness gives all hospitals a score of 3 or 4 on a 4-point scale. The threshold is then set in the transformation so that all hospitals scoring over 3 will have achieved the target, hence variation is reduced to zero. It is unlikely in practice that there is zero variation across England in hospital cleanliness and since it is a performance indicator which is deemed important to include, particularly from the patients' perspective, it may be important to find a different way to measure it. If the rating for hospital cleanliness were instead on a (say) 10-point scale and different thresholds were chosen in the transformation, the indicator might highlight more variation in cleanliness.

Dichotomous variables are similar to threshold variables except they strictly take the form of a yes/no response or pass/fail. These are used in the CPA approach of local government performance assessment. This type of indicator often suffers from the same problems as threshold variables and may display little variation. They may not be sensitive enough to detect differential performance, yet could have a potentially large impact on organisations' ratings in the composite.

Indicators that can be treated as continuous variables are the most common type of format and may include such measures as counts of events, percentages, rating scales, and various transformations such as performance relative to other units. Continuous variables tend to display more variability than other distributions of performance indicators.

Indicators are often published with confidence intervals which is useful if they enable consumers of the information to more readily discern significant differences in performance at the chosen significance level. This allows the important distinction to be made between sampling error and natural variability that would occur within a distribution anyway and true differences in performance for which managers may be held responsible (Jensen, 2000; Street, 2003).

In the NHS star ratings, a number of the clinical indicator variables are published with confidence intervals which are subsequently used to transform the variables into categorical variables (for instance performance above the average for England would be considered one category). Confidence intervals show whether the observation lies in the top or bottom percentiles. Variables such as readmissions are assumed to be drawn from a Poisson distribution, typical of count data. The Poisson distribution assumes that for any given count of events the variance is known and therefore the events are only subject to random variation, but as is explored in section 6 of this report, other sources of variation are likely to be important, which will in general increase the variance. The inability to distinguish these different sources of variation in the clinical indicators, may result in very wide confidence intervals and an inability to distinguish good from bad performance. In reality, the true population value of the mean for mortality rates in a deprived area may be higher than that for less deprived areas, however, the method assumes a single Poisson distribution for England. The actual variation in mortality and readmissions may therefore be higher than is allowed for by the Poisson distribution. Being able to disentangle and decompose the different sources of variation in the clinical indicators may make the use of confidence intervals much more useful since much tighter intervals can be obtained if some of the variation can be reduced.

Change indicators are relevant when longitudinal data is available. In the US study on Medicare patients, the data was collected for 2 periods which enabled the researchers to calculate both absolute improvement on each indicator (defined as the percentage improvement from baseline to follow-up) and relative improvement (defined as the absolute improvement divided by the difference between baseline and perfect performance (100%)) (Jencks *et al*, 2000; Jencks *et al*, 2003). Once again confidence intervals can be calculated for these indicators to assess whether the percentage change is significant or not. Change indicators compare organisational units to themselves rather than to other units. This type of measurement has the disadvantage that units performing poorly in both years may in fact appear better than units performing well in both years, depending on their marginal improvement and thus their incentive to improve year on year may be weak. Year on year changes are also subject to random variation and a higher rate of change in one year is likely to be followed by a lower one in the following year simply due to random variation and the regression to the mean phenomenon, rather that due to actual changes in performance.

5.3.3. Data reliability

As noted earlier, the choice of indicators is often driven by the availability of data, which can result in the exclusion of indicators that are important, but for which data are unavailable or costly to collect. The more comprehensive the composite aims to be, the more likely it is that elements of data will be unavailable in some of the organisations. Similarly, the more diverse the units of assessment are, the greater the chance that the availability of data will also vary. In some circumstances there may be an issue of trading off the relevance of data against its reliability where the importance of including a particular performance dimension may be considered so great that its lack of reliability is considered an acceptable trade-off. This may be if its inclusion at a particular point in time is imperative and more reliable data cannot be collected within the timeframe. This again links to the choice of potentially more timely process measures versus outcome measures with a longer time horizon.

Whatever indicators are chosen, data are unlikely to be 100% complete so the appropriate method for dealing with missing data will always need to be addressed. In the construction of some composite indicators that have wide coverage, units with a lot of missing data could be excluded altogether. In attempting to measure progress towards environmental sustainability across the world (World Economic Forum, 2002), over 50 countries were excluded from the analysis due to limited data coverage, and a number of critical environmental indicators were either not measured at all or were measured imperfectly. Not a single country had measures available for all 68 indicators and the median country included in the index had 16 variables missing. Similarly, the United Nations Development Index, despite using relatively simple dimensions, had to exclude 18 countries due to lack of data (United Nations, 2003). One of the features of the health system performance index created by WHO that has received most criticism relates to the extent of missing data in the construction of the indices. Data were unavailable from 70% of the countries to construct the index of health inequality; in 84% of countries for the two measures of responsiveness; and in 89% of the countries for the fairness in financial contribution (Almeida *et al*, 2001). Thus, as Williams (2001) notes, even for the USA (where we would expect relatively good availability of data), the only indicator that is not imputed is child mortality.

If a decision is taken to impute missing values, it is important that the methods for doing so are made explicit. A literature on methods of imputing missing values exists (Little and Rubin, 1987) including use of mean substitution, correlation results, time series projections or various forms of regression technique, depending on the assumptions made about the nature of the missing values. In creating the indicators of environmental sustainability, a great deal of attention was paid to the most appropriate methods for imputing values and an appendix was produced comparing results from different methods (Abayomi *et al*, 2002). However, the methods used by the WHO in creating their index have been criticised on two grounds: first, they did not make them explicit so readers were not immediately aware of how the imputation had been undertaken; and second, the methods were not "well-documented, validated methods that have withstood peer review" (Almeida *et al*, 2001). One major problem was that much of the actual data on which estimates were based came from a very small (35) sub-set of countries that were not representative, since most were developing countries. So for example, responsiveness was estimated for all other countries on the basis of variables found to be predictive of responsiveness in the 35 countries surveyed. However, as Nord notes (2002) it is far from evident that relationships between dependent and predictive variables are the same in developed and developing countries. So the imputed score for the USA on responsiveness is high in part because several responsiveness variables were found to be related to income in the 35 countries surveyed.

5.3.4. Collinearity of indicators

There will often be high correlations between certain performance indicators which are measuring similar aspects of performance. The concern is that, in the construction of a composite, the inclusion of variables which are highly collinear will effectively introduce some sort of double counting. It has therefore been argued that inclusion of a reduced set of indicators based on a choice between those indicators with high correlations may be desirable for reasons such as parsimony and transparency. Multivariate statistical methods to investigate relationships between the indicators of a composite include Principal Components Analysis (PCA) and Factor Analysis (FA). These methods may be used to extract statistical correlations between indicators enabling a core group of indicators to be identified which statistically best represent the remaining excluded indicators (Joint Research Centre, 2002).

However, it is not strictly necessary from a technical point of view that highly collinear variables be excluded. For instance, if two perfectly collinear variables were included in the composite, with weights w_1 and w_2, then the particular dimension of performance which they measure will be included in the composite with the weight

(w_1 + w_2). This is not problematic if the weights have been chosen correctly (Smith, 2002).

If statistical techniques are used to choose the variables for inclusion, then it is likely that variables may be chosen on the grounds of statistical significance. In this case it is likely that highly collinear variables will be excluded through model specification tests for multicollinearity. However, if there is a high degree of collinearity between the indicators for potential inclusion in the composite, then the model selection procedures used may be highly influential in determining the chosen indicators and therefore somewhat arbitrary too. The choice of one variable over an alternative highly collinear variable, may not alter rankings greatly, but may affect the judgements on a small number of units with extraordinary performance in either of those dimensions. It may therefore be subject to dispute and challenge.

In studies examining whether a single indicator may perform as well as a composite indicator, male unemployment levels and car ownership were found to explain nearly as much variation in health status as various composites of deprivation (Ellaway, 1997). Similarly, application of principal component analysis to the composite indicator of Human Development produced by the United Nations suggested that just one dimension of the composite (life expectancy) could be used without loss of too much information (Joint Research Centre, 2002).

In practice, initiatives to improve performance are likely to impact on more than one performance indicator, thus performance indicators cannot truly be considered independent and collinearity is an inevitable consequence of such relationships (Smith, 2002).

5.3.5. Level of data aggregation

The level of data aggregation reflects important decisions around whether the chosen indicators are aggregated enough to be comprehensive, but disaggregated enough to be detailed. In the NHS star rating system, a number of the indicators are not in fact applicable to all units being assessed. This may occur for a number of reasons, for instance, the indicator may measure activity which is not relevant to the particular hospital, or the numbers may just be too small (such as waiting times for heart patients, or CABG death rates). A higher level of data aggregation will therefore ensure all hospitals can be measured on the chosen indicator, however this may lead to the exclusion of important disaggregated indicators.

Careful consideration therefore needs to be given to the choice of variables (and the weight which is attached to them) which only measure performance relevant to a sub-sample of the units. Such incomplete coverage due to a lower level of data aggregation, may suggest the need to rather produce composite indices for sub-groups of units for which comparable relevant indicators can be found.

5.4. Transforming the indicators

There is no need for any transformation if it is possible to specify a weight that indicates the relative value to the composite of an extra unit of attainment in that dimension *at all levels of attainment*. Otherwise a transformation is required. The objective is to make the transformed variable such that an extra unit of attainment is of equal value at all levels of attainment.

In most cases the indicators will be measured in different units which will reflect different weights in the composite and therefore need to be transformed or standardised in some way before they can be aggregated together in a composite. Variables are transformed to a common basis to avoid problems of mixing different units of measurement (such as labour, expenditure, events). Variables are also normalised to avoid problems with outliers or extreme values. If variables have very skewed underlying distributions they can be normalised through logarithmic transformations or they could be truncated to exclude extreme outliers (Freudenberg, 2003). There may be arguments for trimming the tails in the raw data in order to avoid extreme values, especially if there is good reason to assume the main cause of the extremes is due to poor data quality. So for example, the data in the Environmental Sustainability Index was truncated at values outside the 97.5 and 2.5 percentile scores (World Economic Forum, 2002).

In the most simple form, assume a composite indicator takes a linear form as follows:

$$CI_n = w_1 y_{1n} + w_2 y_{2n} + \ldots + w_p y_{pn} = \sum_i w_i y_{in} \tag{2}$$

where CI_n is the composite score for unit n, y_{in} is the individual performance measure for attribute i in unit n, and w_i is the weight attached to attribute i (Smith, 2002).

Thus the relative price of the outputs, or the ratio w_1/w_2 which indicates the amount of objective 2 that needs to be sacrificed in order to gain an extra unit of objective 1, needs to be constant, regardless of the observed values of the indicators. The purpose of the transformation is to ensure that such invariance is secured.

The statistical distribution of the underlying variables is immaterial to the specification of the composite index. However what is important is that, assuming a linear composite, transformation of the variables may ensure that the weights used in the composite are valid across the range of observed performance (Smith, 2002). In other words, the transformations that are applied will have an effect on the interpretation of the weights in the construction of the composite. This issue is highlighted in the following section outlining the methods that exist for transforming and normalising the indicators used in constructing a composite (Joint Research Centre, 2002; Freudenberg, 2003).

5.4.1. Rankings

This simply ranks each unit on each of the indicators as follows:

$$y_{in} = Rank(x_{in}) \tag{3}$$

where *y* is the transformed variable of *x* for indicator *i* for unit *n*.

The composite indicator is then created by either a sum of the rankings or an average of the rankings.

Ranking is therefore based on ordinal levels so the main disadvantage is the loss of absolute level information. It does not allow conclusions to be drawn about the difference in performance between the units being assessed as there is no measure of the distance between the values of the indicator.

The sum of rankings method has been used to create a composite measuring the development and application of information and communication technology across countries, by aggregating rankings on 5 simple indicators (Fagerberg, 2001), while the use of average rankings was one of the methods used in the Medicare study on healthcare performance across US states (Jencks *et al*, 2000; Jencks *et al*, 2003).

5.4.2. Normalising or standardising to a z-score

This method imposes a standard normal distribution onto each indicator and is the method used in section 6 of this report. Each indicator will therefore have a mean of zero and a standard deviation of 1 (N~(0,1)). The formula for calculating this is:

$$y_{in} = \left(\frac{x_{in} - \bar{x}_{in}}{\sigma_{\bar{x}_{in}}} \right) \qquad (4)$$

where $\bar{x}$ is the mean value and σ is the standard deviation. Thus it converts all indicators to a common scale in which they are assumed to have a normal distribution. All indicators will therefore have a mean of zero. Whilst standardised scores for each indicator deals with outliers to some degree, it still allows extreme values to influence the results because the range between the minimum and maximum standardised scores will vary for each indicator – thus, it gives greater weight to an indicator in those units with extreme values. However, this may be desirable if the intention is to reward exceptional behaviour – if an extremely good result on a few indicators is thought to be better than a lot of average scores (Joint Research Centre, 2002).

This approach has been used in the construction of many composite indicators such as a composite of investment in the knowledge based economy or a country's capacity to create knowledge (Muldur, 2001) and the environmental sustainability index (World Economic Forum, 2002). The construction of the WHO index of health system performance has been criticised for failing to use appropriate methods for transforming the data and the use of z scores has been recommended for future work (SPRG, 2001).

5.4.3. Re-scaled values

Often before applying the method of standardisation to a z-score, re-scaled values are created in order to give an identical range for the standardised scores for every indicator. Re-scaling ensures that the transformed indicators are given a value relative to the global maximum and minimum and the re-scaled index takes a value from 0 (worst) to 100 (best) as follows:

$$y_{in} = \left(\frac{x_{in} - \min(x_i)}{\max(x_i) - \min(x_i)} \right) \times 100 \tag{5}$$

Here standardisation is based on the range rather than the standard deviation and these extreme values (the minimum and maximum) may be unreliable outliers. Whilst the method may be more robust where there are lots of outliers, the range for indicators with very little variation will increase and these will contribute more to the composite indicator than they would using the un-scaled method. This method is therefore more dependent on the value of the weightings for each indicator than the un-scaled method where the contribution of each indicator to the composite depends on both the weighting and the variance of the indicator. Thus the re-scaling method is linked with the issue of the choice of weights.

5.4.4. Ratio/percentage difference from the mean

This method takes the ratio or percentage distance from the mean for all units for each indicator as follows:

$$y_{in} = \left(\frac{x_{in}}{\bar{x}_{in}} \right) \times 100 \tag{6}$$

Thus the (weighted or unweighted) mean is given a value of 100 and units receive a score depending on their distance from the mean. Values greater (less) than 100 indicate above (below) average performance. Statistically, however, it is less robust to the influence of outliers than some of the other methods.

5.4.5. Percentage difference from leader

This method assigns the leader unit on the particular indicator a value of 100 and all other units are then ranked as a percentage away from the leader as follows:

$$y_{in} = \left(\frac{x_{in}}{\max(x_i)} \right) \times 100 \tag{7}$$

5.4.6. Number of indicators above the mean minus number below the mean

This method defines an arbitrary threshold around the mean and takes the difference between the number of indicators above and below the mean. The main drawback is the loss of interval level information as units will be assigned as being above/below average regardless of how much better/worse they are. This may be applied as follows:

$$y_{in} = \frac{x_{in}}{\bar{x}_{in}} - (1 + p) \tag{8}$$

where *p* is an arbitrary threshold above and below the mean.

A summary innovation index for European countries, using 17 indicators across 4 areas used this approach (European Commission, 2001a). For each indicator, the threshold was defined as 20% of the European average value for the indicator and the index for a country is equal to the number of indicators that are more than 20% above the overall mean, minus the number of indicators that are more than 20% below it.

The 20% threshold was arbitrary but allowed for non-meaningful differences from the mean that may have been attributed to data errors or variation in definitions. A sensitivity analysis showed a high correlation between the index produced with the 20% threshold and those produced using 15% and 25% thresholds.

The advantage of this method is that it is more robust to outlier data than several other methods. However, the method loses interval level data since, for instance in the example above, countries A and B with scores of 300% and 25% above the mean respectively on indicator x with threshold *p* = 20% will be considered equally above average.

5.4.7. Percentage of annual differences over time

The approach taken in this method is to use the values of the indicator from previous years to standardise as follows:

$$y_{in}^{t} = \frac{x_{in}^{t} - x_{in}^{t-1}}{x_{in}^{t}} \tag{9}$$

where *t* indexes time. The value assigned to each indicator is the difference in the value between the current year and the previous year, divided by the value at the previous year. This approach will of course only be feasible where longitudinal data is available and the indicators are measured in the same way over time. This method effectively smoothes against cyclical variability.

An index attempting to measure the extent to which potential benefits of the European internal market have been realised, was constructed using this approach (European Commission, 2001b).

5.4.8. Logarithmic transformation

A common transformation is to use logarithms of all indicators as follows:

$$y_{in} = \ln(x_{in}) \tag{10}$$

In this case, we would re-interpret the coefficients from equation (2) as elasticities. Thus the ratio w_1/w_2 now indicates the percentage improvement in objective 1 that would compensate for a one percentage point decline in objective 2. In this case the marginal utility of an improvement in an indicator declines as performance improves. Thus the transformation will lead to a higher weighting for a one unit improvement from a low level of performance compared to an identical one unit improvement from a high level performance.

The impact on health outcome of a one percentage point improvement on two different indicators may vary substantially across patient groups. For example, a one percent improvement in aspirin prescribed at discharge may have a very different impact on health outcomes for AMI patients than would a one percentage point improvement in smoking cessation counselling. The incentive will be to concentrate on areas where the costs associated with securing improved rankings are lowest (Smith, 2002).

5.4.9. Transformations to a categorical scale

This is where each indicator is assigned a categorical score. This may be either numerical, such as the three or five point scale used in the star ratings, or it may be qualitative, such as ‘achieved’, under achieved’ and ‘significantly under achieved’, another approach also taken in the star ratings.

Thresholds need to be chosen to assign scores to different categories. Categorical scales tend to be highly subjective as they depend largely on the choice of thresholds which may be selected arbitrarily. Categorical scales also omit a great deal of information about the variance between units on the transformed indicators.

One method of choosing thresholds is to base it on percentiles of the distribution. Thus for instance the top 10% of units receive a score of 1, the next 20% receive a score of 2, the next 20% receive a score of 3 and so on up to the last 10% which receive a score of 7. Thus organisations are compared against one another rather than against a standard. This is a common approach for many of the indicators used in the CPA for local authorities.

One disadvantage of this approach is that even if there is little variation within the original scores, the percentile banding will force this categorisation onto the data, irrespective of the distribution of the underlying data. This is particularly problematic if the data is highly skewed. In many examples, including the CPA and the star ratings, the institutions responsible for the construction of composite indicators, have varied the percentile brackets across different individual indicators with different underlying distributions in order to obtain transformed categorical variables with more ‘normal’ distributions.

In terms of year on year analyses using percentile transformed data, there is the advantage that any small changes in the way the indicator is measured will not affect the transformed variable if the same percentile transformation applies. For instance, student exam scores transformed to percentiles will ensure that a certain percentage of students pass each year even if the difficulty level changes. On the other hand, the disadvantage of this transformation is that overall improvements year on year will not be picked up.

Examples of the above transformations are shown in the following table with some hypothetical data for 10 organisations to illustrate the different approaches and how sensitive the results can be to the choice of transformation technique. Potential problems include the loss of interval level information (e.g. ranking, categorical scale, and threshold above and below mean), sensitivity to outliers (e.g. standardising, difference from mean, and difference from leader), arbitrary choice of thresholds (e.g. threshold above and below mean, and categorical scale), and sensitivity to weighting (e.g. re-scaling, and logarithmic transformation). These different transformations will therefore have important effects on the construction of the composite indicator, and important incentive effects on the behaviour of units being assessed.

Table 16: Examples of different transformation methods

Unit	Raw data	Ranking	Standardising	Re-scaling (Best = 100, worst = 0)	Difference from mean (Mean = 100)	Difference from leader (Best = 100)	Threshold above and below mean (Threshold = 20%)	Logarithmic	Categorical scale (Percentiles = 0.75, 0.5 and 0.25)
Unit 1	2.85	1	2.01	100	174	100	1.37	0.45	3
Unit 2	2.05	2	1.09	71.01	94	71.93	0.65	0.31	3
Unit 3	1.58	3	0.54	53.99	47	55.44	0.22	0.20	3
Unit 4	1.35	4	0.28	45.65	24	47.37	0.02	0.13	2
Unit 5	1.03	5	-0.09	34.06	-8	36.14	-0.27	0.01	2
Unit 6	0.86	6	-0.29	27.90	-25	30.18	-0.43	-0.07	1
Unit 7	0.59	7	-0.60	18.12	-52	20.70	-0.67	-0.23	1
Unit 8	0.43	8	-0.79	12.32	-68	15.09	-0.81	-0.37	0
Unit 9	0.28	9	-0.96	6.88	-83	9.82	-0.95	-0.55	0
Unit 10	0.09	10	-1.18	0	-102	3.16	-1.12	-1.05	0
Unweighted average = 1.11 Standard deviation = 0.86									

5.4.10. *Economic considerations in transformations*

As mentioned, the process of the transformation of indicators is linked to the interpretation of the weights attached to the indicators and is therefore crucial in terms of the incentives which may be generated by the implicit weights. Furthermore, these transformation methods are also associated with different incentive effects for the organisations being assessed.

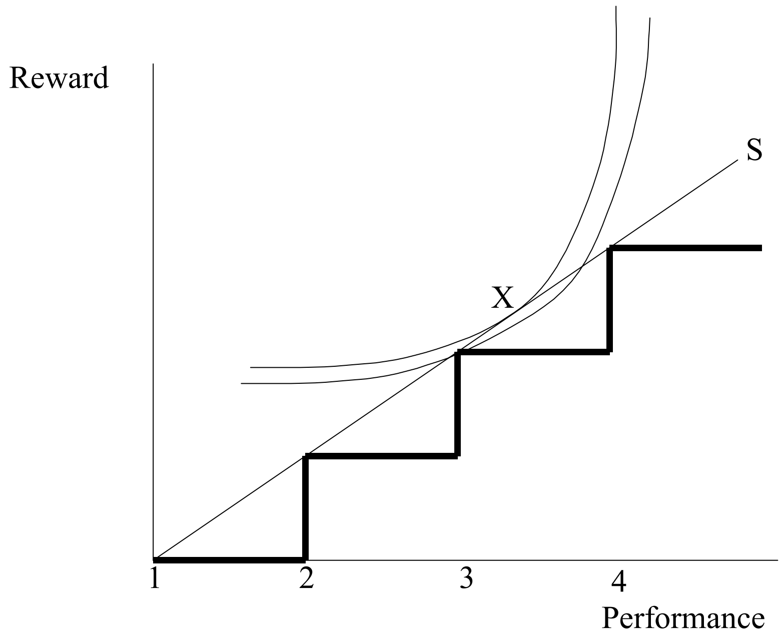
When variables are transformed onto a categorical scale, some basic economic issues arise. Using discrete choice theory and assuming rational managers (and no uncertainty or measurement error), economic theory predicts that only the thresholds will be chosen as production points (as in Figure 2).

From the regulator's viewpoint, the main rationale for introducing such step functions (compared to a continuous reward function) could be that

- it induces (on average) improved performance, or
- it reduces the rewards that must be paid, or
- it reduces disparities between organizations (for example encouraging certain minimum standards).

Figure 2 shows the original flat reward schedule S when a unit is producing one output (one indicator) which is measured on a continuous scale.

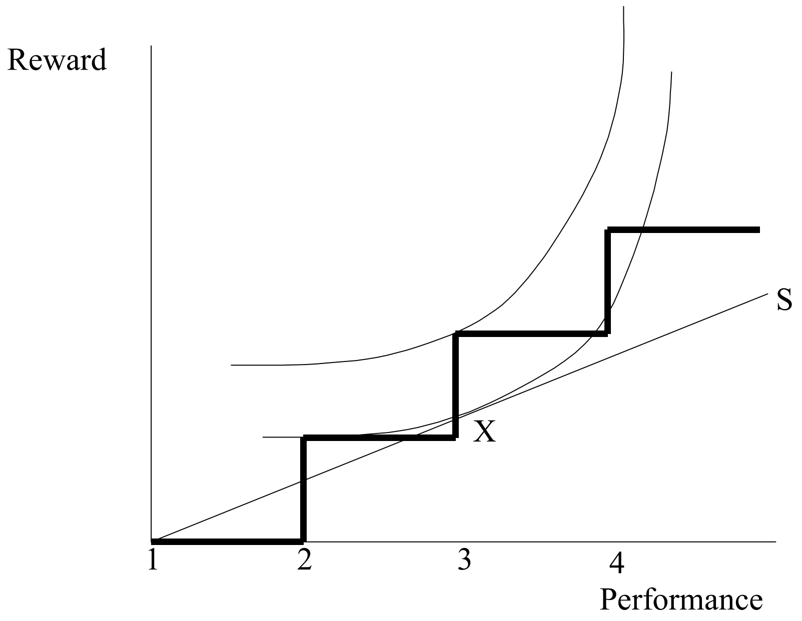
Figure 2: Link between performance and reward, one output



In Figure 2, with just one output, the original performance reward schedule is converted into the step function in bold. (Note that the indifference curves reflect an implicit utility function based on *performance* and reward, where the underlying utility function should be based on *effort* and reward (and possibly performance if there is a degree of managerial altruism). In general, this will change the preferred production point, in this case a reduction from point X to step 3.

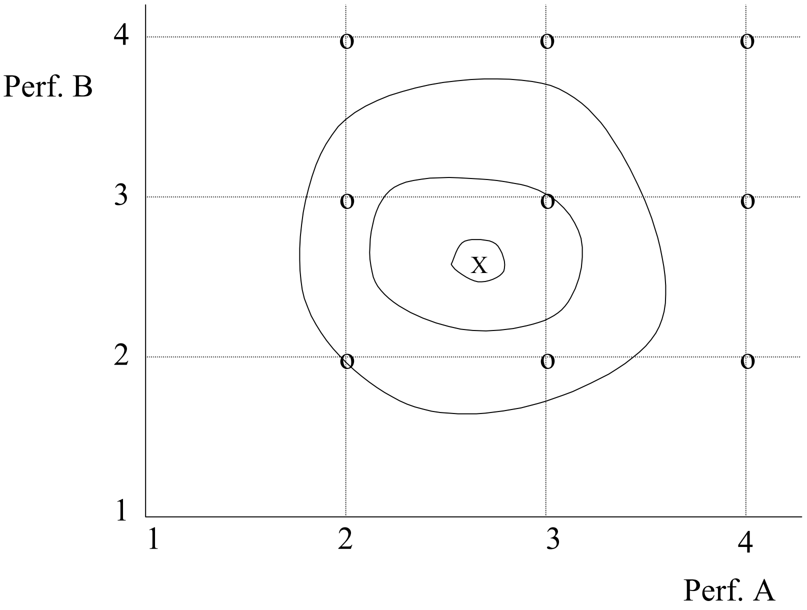
Figure 2 assumes the reward schedule is unchanged in power from the original schedule S. It is quite likely that a more powerful reward schedule will be put in place by the introduction of the step function. This is shown in Figure 3, which shows that – even when the reward schedule becomes steeper, there might still be a diminution of performance from point X to step 3.

Figure 3: Performance and reward, steeper reward schedule



Equally, introducing two outputs can lead to changes in performance up or down, as illustrated in Figure 4. Given the current reward schedule, the optimal production point with continuous variables is X. If the power of the schedule remains unchanged, introduction of step functions leads to a lattice of potential optimal production choices, and the optimal choice depends on the preference map (in this case, the choice will be (3,3)).

Figure 4: Production choices under step functions, two outputs



Hence, the transformations of indicators to categorical variables, may have important incentive effects and the choice of thresholds will play an important role in how strong those incentives are. The fact that this transformation is particularly widely used in the public sector in the UK in the construction of composites suggests a careful scrutiny is required of the choice of thresholds since these can be very subjective, particularly if they change from one indicator to the next in order to secure approximately 'normal' distributions on all transformed variables.

5.5. Combining the indicators

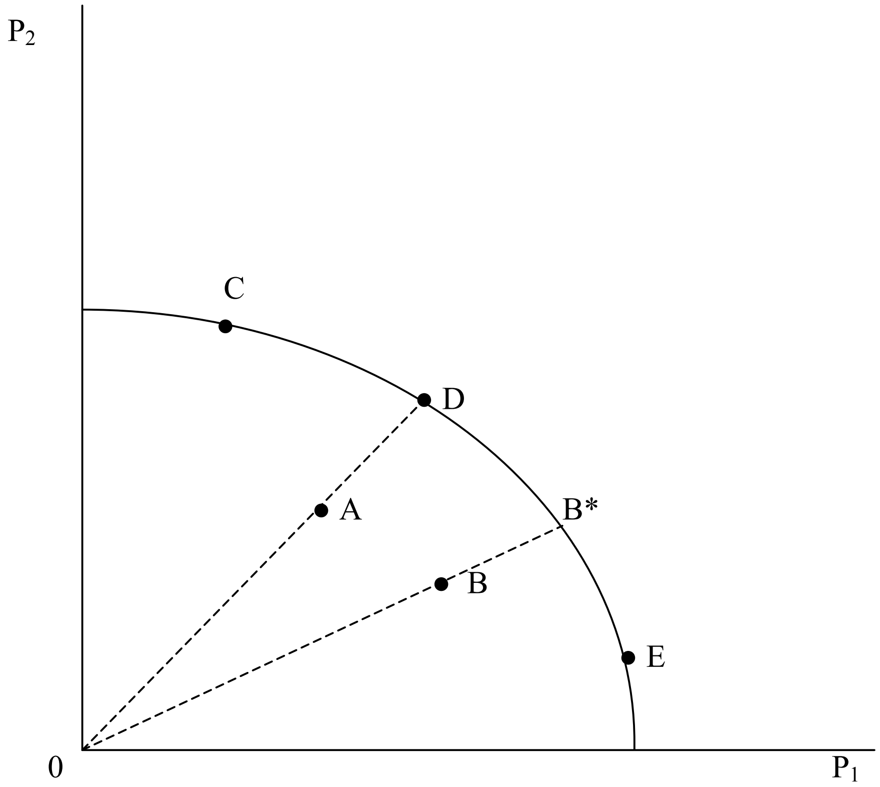
The different dimensions of performance measured on different scales (which are then transformed into a common scale) then need to be combined in a meaningful way. This gives rise to some questions, namely what weights will be applied to the individual indicators, whose preferences will these weights reflect and how will they be elicited, and finally, what decision rules will be applied to combine the transformed indicators into a composite?

5.5.1. Weighting

When variables are aggregated into a composite they need to be weighted in some way. All variables may be given equal weight or they may be given different weights which reflect the priority, reliability or other characteristics of the underlying indicators (Freudenberg, 2003). As mentioned, weights may also be chosen to reflect a constant value for the relative price of the outputs (if variables are not transformed), although this may be difficult often to accomplish in practice.

Figure 5 shows the observed performance of five units with identical expenditure and environmental circumstances. This is depicted using a production possibility frontier which shows the maximum technologically feasible level of performance on outcome (indicator) P_2 that can be obtained for every feasible level of outcome P_1. For the chosen level of expenditure, the units obtain the observed mix of the two performance outcome measures P_1 and P_2. In this example, one can unambiguously state that unit D's performance is preferred to unit A which is technically inefficient. It lies inside the efficient production possibility frontier. Also system B is inferior to a linear combination of systems D and E, represented by point B*. However, no "objective" ranking of units C, D and E can be made without a judgement on the relative weights to attach to outcomes P_1 and P_2. These cannot be ranked without reflecting preferences for outcome P_1 relative to outcome P_2. As the expenditure on the health system increases, the production possibility frontier will expand outwards, and given variable returns to scale in production this expansion is likely to not be symmetric. At different levels of activity, improvement in some outcomes may be easier than in others. Furthermore, over time, as the production frontier expands, it is unlikely that relative weights will remain constant. Thus at very low levels of expenditure, certain outcomes might be weighted higher than when expenditure increases (Smith, 2002).

Figure 5: Observed performance of five units with identical expenditure and environment



The intention of weights is therefore to indicate the relative importance of indicators. As shown from equation 2 in the previous section, the ratio w_1/w_2 indicates the amount of objective 2 that needs to be sacrificed in order to gain an extra unit of objective 1. Thus the weights are analogous to prices for the outputs.

Weights are essentially value judgements about the relative importance of different performance indicators and about the relative opportunity cost of achieving those performance measures. The rationale for the application of differential weights is that greater organisational effort will be used to achieve better performance on those dimensions which are considered more important. Weighting adds the additional component of trying to induce behaviour change, since it makes explicit the objective function underlying the construction of the composite (tying it back to the second step in the construction of the composite indicator).

The weights which are attached to different performance indicators have a profound effect on the outcome of the composite index and can dramatically change the ranking of a particular unit if an indicator is given more weight on which the unit either excels or fails. The weighting system which is applied (whether using statistical methods, an underlying theoretical framework, or a conceptual rationale) needs to made explicit and transparent. There is also scope to apply sensitivity analysis to assess the robustness of rankings to the application of different weighting systems.

Weights may be applied to reflect the underlying data quality of the indicators thus giving less weight to those variables where data problems exist or with large amounts of missing values. The reliability of a composite can potentially be improved if it gives more weight to good quality data. However, this may as a result give more emphasis to indicators which are simply more easy to measure and readily available rather than more important measures which may be more problematic to identify with good data.

Often equal weights are applied to all underlying indicators, simply for simplicity sake, which suggests that all indicators have equal importance in the composite. This is the explicit approach taken for the majority of the domains in the policing performance assessment. There is still of course a judgement being made in terms of the relative importance of each indicator, but if there are no statistical or empirical grounds for choosing different weights, this may be a valid approach in some contexts. For example, the environmental sustainability index used equal weighting as the authors felt there was no firm basis for applying differential weights given the existing knowledge and no scientific consensus existed (World Economic Forum, 2002). Similarly the Summary Innovation Index (European Commission, 2001a) used equal weights to combine the indicators as the authors felt there was no generally applicable model to explain how each indicator influenced innovation.

Some commentators have argued that the equal weights approach may give extra weight to certain performance aspects if several underlying indicators are in effect measuring the same attribute (Freudenberg, 2003). This may not however, be a concern, since high correlations may simply show that performance on these indicators by a particular unit is similar (which is indeed very likely). For instance, if hospitals perform well on certain indicators of waiting times, it is likely that they will also perform well on others. The correlations between indicators is therefore closely linked to the issue of weights. It has been argued that equal weights are appropriate either where the indicators within the composite are uncorrelated or are highly correlated, but less so where some are correlated and some are not (Joint Research Centre, 2002).

The following equations show the relationship between the weights and the correlation structure between indicators. From equation (2), the composite C takes the following form:

$$C = \sum_{i=1}^{n} w_i x_i \tag{11}$$

where *n* is the number of indicators, w_i is the weight attached to indicator *i*, and x_i the score on indicator *i*.

The variance of this composite is calculated as:

$$\sigma_C^2 = \sum_{i=1}^{n} w_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} w_i w_j \operatorname{cov}(x_i, x_j) \tag{12}$$

leading to the following:

$$\frac{\partial \sigma_C^2}{\partial w_i} = 2 w_i \sigma_i^2 + 2 \sum_{j \neq i} w_j \operatorname{cov}(x_i, x_j) \tag{13}$$

So an exogenous increase in weight w_i leads to an *increase* in variance, providing $w_i \sigma_i^2 + \sum_{j \neq i} w_j \operatorname{cov}(x_i, x_j) > 0$.

Usually this will be the case, as ‘good’ organisations tend to score well across most indicators (that is, cov(x_i,x_j)>0 for most (*i,j*)).

It might be argued that we are only interested in increases in the *relative* weight attached to indicator *i*. This is trivial to accommodate, by dividing each weight by a scaling factor:

$$W = \sum_{i=1}^{n} w_i \tag{14}$$

Then:

$$\sigma_C^2 = \left\{ \sum_{i=1}^{n} w_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} w_i w_j \operatorname{cov}(x_i, x_j) \right\} / W^2 \tag{15}$$

and

$$\frac{\partial \sigma_C^2}{\partial w_i} = \frac{2}{W^2} \left\{ w_i \sigma_i^2 + \sum_{j \neq i} w_j \operatorname{cov}(x_i, x_j) \right\} \left(1 - \frac{w_i}{W} \right), \tag{16}$$

yielding broadly similar conclusions.

There are several methods available to extract correlations among indicators. Notably, Principal Components Analysis (PCA) and Factor Analysis (FA) could be used to define weights for indicators within the composite based on correlations between the indicators. The weights fall out of the statistical methods used to investigate the relationships between the indicators since the methods essentially reduce the dimensionality of the data down to a small number of principal components or factors which measure unique "statistical" dimensions in the data. The disadvantage of using this approach to weighting is that the correlations do not necessarily correspond to the real-world links and underlying relationships between the indicators and the phenomena being measured (Joint Research Centre, 2002). PCA has been used to construct composite indicators in several applications including the internal market index in EU countries which was based on large sets of underlying indicators (European Commission, 2001b).

In other circumstances, the PCA approach to weighting has been rejected as inappropriate: for example, in constructing an index of composite leading indicators for the OECD (to measure business climate for forecasting purposes), equal weights were used for indicators as the authors claimed PCA would minimise the contribution of indicators that did not move with the other indicators. This was significant as there was variation in how well indicators performed between cycles (European Commission, 2000). In constructing the index of environmental sustainability, equal weights were used in preference to those suggested by PCA as the latter assigned negative weights to some indicators (World Economic Forum, 2002).

Another interesting approach that has been used in some contexts is the "distance to target" approach (Joint Research Centre, 2002). In this approach, the weights are determined by dividing the indicator values by a target value (expressed in the same units) and then aggregating across all indicators. The idea is that the weights reflect the urgency of the problem and the relative need for policy attention – which is high for units that are distant from the target, and low for those nearer the target. This approach works best where there are clearly defined national goals and therefore is more suitable for within-country comparisons. It has been used in the Netherlands to measure whether environmental policy was heading in the right direction by applying the targets for the indicators to the observed values over time (for instance targets for noise disturbance, acidification, dispersion of toxic substances and so forth) (Adriaanse, 1993). The origin of the targets is clearly an important issue as the indicator will only be acceptable if the targets are viewed as appropriate. It is also a relatively narrow approach as it assesses the benefits or outcomes of the policy only in terms of the specific policy goals rather than taking account of wider considerations.

The efficiency frontier approach is another possible way in which the weights can be applied to the data in so far as the weights are actually determined by the data, depending on where the unit of assessment is located relative to those units on the efficiency frontier. This has been used to construct an index of labour market performance in Europe, combining 3 measures of unemployment (Storrie and Bjurek, 2000). However, one criticism of such an approach is that it is does not give an immediately obvious indication to each unit of what direction they should move in, in order to improve performance.

5.5.2. Eliciting preferences

The above methods of PCA and FA and the efficiency frontier approach are all essentially statistical or analytical techniques used to obtain relevant weights for the underlying indicators. However indicators could also be weighted using the judgement of individuals based on survey methods to elicit their preferences. There are of course fundamental considerations around whose preferences will be used in the application of those weights (i.e. whose objective function is to be made explicit), whether it be the preferences of policymakers, providers, purchasers, patients, or the public, and furthermore, how the preferences of those individuals (or groups of individuals) will be elicited.

The weights used reflect a single set of preferences, whilst the evidence suggests there exist a great diversity in preferences across policy makers, individual unit actors and the broader public. There is likely to be considerable variation in the preferences of respondents. This should temper the notion of presenting the composite as "objective" (Smith, 2002).

To date, there has been very little consideration for public preferences, even though evidence suggests these may differ greatly from policymakers' preferences. If the NHS is to become more consumer orientated, then it needs to find a way of incorporating public opinions about priorities. The market is the usual economic process by which consumers bring their value judgements and preferences to bear on their individual resource allocation and consumption decisions. However market failures in health care make it difficult for preferences to be revealed, thus requiring alternative ways to obtain the relative weights. This section describes the construction of various composites where different groups of individuals have been surveyed to obtain their preferences.

Where the dimensions of a composite are very technical in nature, the use of "expert" opinion has been advocated - for example, in assessing various sorts of emissions and their impact on the environment - this may prove difficult for individuals with no background in the topic. However, one difficulty is deciding on what sort of experts should be asked and how many should be included. There may be a wide range of views about certain policy issues and it is therefore important that the group is not dominated by experts with a very specific set of interests or a limited outlook. A great deal of attention has been paid to the selection of experts in some instances. For example, in order to create an index of "environmental friendliness", weights were obtained from a survey of respondents from 8 sectors, thought to represent a variety of expert stakeholder perspectives (such as environmental scientists, environmental journalists, representatives from agriculture, manufacturing and so on) (Eurostat, 2000). The valuations varied between groups, and the authors note that respondents were especially conscious of their "own" problems, so for example, those working in the traffic sector weighted "noise" more heavily than others.

The weights for the 3 indicator categories in a composite designed in the Netherlands to measure environmental impact in Europe were also chosen based on expert opinion (PRé (Product Econology Consultants), 2000). Although a large postal survey of experts from a group of Swiss experts was undertaken, the final weights were based on only 45 responses. Guided by responses about ranking and weighting and supplementary questions, the respondents were divided into 3 cultural “types” that varied in terms of their outlook on environmental issues and the authors found the weights varied significantly with cultural type. They present the weights by each cultural type (although this was based on only 29 respondents) and their subsequent analysis is presented in terms of the “default” (average) weights and the weights from each cultural type.

The group of respondents used to derive the weights for summing the 5 indicators of the WHO index were likely to have very specific views as over half were WHO members of staff and almost all were professionals working in the health sector. This approach has therefore been heavily criticised (Williams, 2001). However, it has since been argued that further investigation of weights from a more representative population showed very similar results (SPRG, 2001), although this still appears to be a matter of debate (Almeida *et al*, 2001). Other studies have however found little variation in results, for example a study to elicit weights for environmental indicators involving 400 experts drawn from a wide range of social spheres such as industry and the environmental sector, found a great deal of consistency in the results (Moldan *et al*, 1997).

In circumstances where the concepts can be easily understood by a non-specialist, the general public rather than experts may be considered an appropriate group from whom to elicit weights. A composite aimed at measuring public concern for environmental issues used the views of the public derived from various opinion polls (Joint Research Centre, 2000). Five different weighting schemes from several different countries and years were tried in the analysis but the results did not vary significantly, leading the authors to conclude that public opinion about the main threats to the environment is stable across time and space and that the public were rational in their evaluations. In the preferences elicited by the public for the six indicators in the Kings Fund study on health authority performance, they concluded that there was little difference in the preferences or weights (“budget chips”) assigned by those from different gender, age, social class group or area (Appleby & Mulligan, 2000). However, while this may be the case on average, it is likely that some heterogeneity would still exist within groups.

5.5.3. Methods of elicitation

Different approaches for eliciting preferences or values include single and multiple voting, simple scoring and scaling, amongst others, as well as more complicated methods such as analytic hierarchy process and conjoint analysis, each with their own advantages and disadvantages (Mullen and Spurgeon, 2000). There appears to be little consensus however as to which is the preferred technique (Dolan *et al*, 1996).

Budget allocation was used in the Kings Fund study on health authority performance (Appleby & Mulligan, 2000) where participants were given a budget to be distributed across the different priorities, thus revealing their preferences for greater expenditure on the indicators of greater importance. A similar approach was used in the study to elicit weights for environmental indicators involving 400 German experts in which they were asked to allocate a budget to several environmental indicators related to air pollution (Moldan *et al*, 1997). The budget allocation approach however cannot readily be applied if there are too many indicators and it is argued the approach is optimal for a maximum number of 10 indicators (Joint Research Centre, 2000).

Public opinion polls have been used to elicit weights in some contexts (such as public concern for environmental issues) and weights have been derived by multiplying the proportion of people choosing a specific response by a score assigned to each response, then converting them to sum to one. However, a degree of weighting is already introduced before the public give their views as decisions are made about the scores. For example, trying to elicit views on the degree of public worry about environmental issues, the responses were scored as 3 for "great deal"; 2 for "fair amount" and 1 for "not very much" (Joint Research Centre, 2002). Thus the authors of the poll have already influenced the subsequent weights.

The Analytic Hierarchy Process (AHP) is a technique for multi-attribute decision-making in which opinions are systematically extracted by pair-wise comparisons (Saaty, 1987). AHP enables decomposition of a problem into a hierarchy and ensures that both the quantitative and qualitative aspects of the problem are incorporated into the evaluation process. Respondents make ordinal pair-wise comparisons of attributes and express their strength of preference on a semantic scale of 1-9 in which 1 indicates equality and 9 indicates that the indicator is 9 times more important. The pair-wise comparisons result in a matrix which reflects the ratios of relative contributions (weights) of the indicators. The advantage of AHP is that pair-wise comparisons are more manageable in decision-making problems than multiple trade-offs. One caution with the AHP approach is that individuals' preferences are not always consistent although the method allows the calculation of an inconsistency ratio which if low does not drastically affect the weights. The Analytic Hierarchy Process has been used to determine weights in the environmental sector (Eurostat, 2000) for 11 components of the environmental friendliness index.

Direct interview techniques have been used to elicit responses on the extent to which the public is prepared to sacrifice health gain for reductions in health inequalities (Shaw *et al*, 2001). This two-year study derived a way of asking questions on various inequality issues which enabled people to indicate their strength of preference for different sorts of reduction in health inequalities and thus enabled members of the general public to make meaningful trade-offs between efficiency and equity.

Conjoint analysis is a survey method developed in mathematical psychology and has gained widespread use in health care (Ryan and Farrar, 2000). The technique is based on the premise that a good or service can be described by its attributes and the extent to which an individual values the good or service depends on the strength of these attributes (characteristics). The technique can therefore be used to elicit weights from the relative importance of different attributes of a service. The method involves drawing up scenarios with all possible service (or outcome) configurations, although all scenarios can rarely be included in a questionnaire and are usually reduced to a manageable number.

Preferences for scenarios are then elicited by either ranking, rating or discrete choice. Rating requires respondents to assign a score on say a five-point scale to each of the scenarios. Discrete choice requires respondents to either state that scenario A or B is preferred or assign a score on a five point scale where 1 equals definitely prefer A and

5 equals definitely prefer B. Regression techniques are then used to analyse individuals' responses with the coefficients showing the relative importance (weights) of the different attributes as well as whether the attribute has a statistically significant effect on the choices. As with AHP, checks can be made for internal consistency as well as dominant preferences, where respondents are not willing to trade a reduction in one attribute for an improvement in another. Conjoint analysis has as yet not readily been used as method for eliciting weights in the construction of composite indices and whilst it holds promise as a rigorous survey technique, it still holds some methodological challenges.

There are also various innovative ways in which the elicited weights can be explored. In the Eco-indicator 99 project the weights were chosen by a panel of 45, pertaining to three damage categories, namely Human health, Ecosystem quality and Resources.

Respondents were asked to rank and weight the three damage categories. The average weights are shown in the table below.

Table 17: Average weights for three damage categories

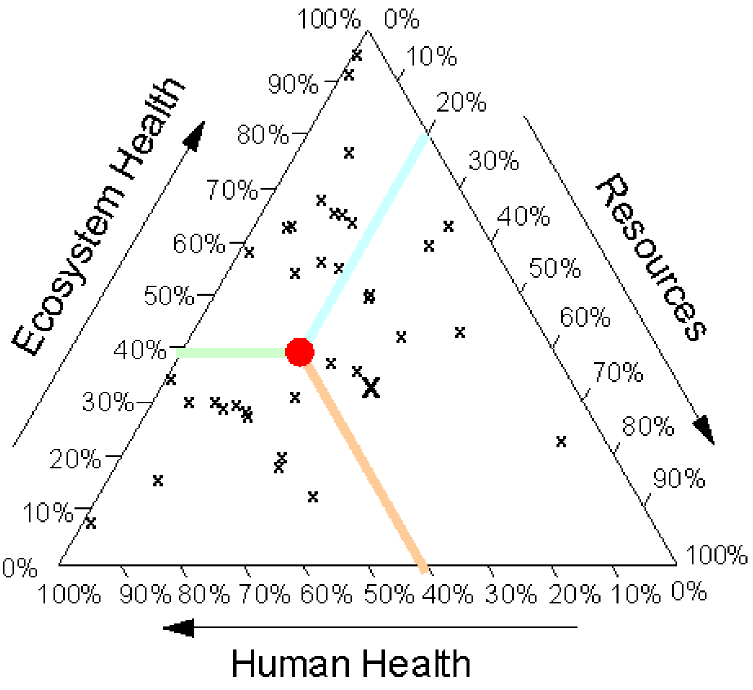
	Mean	Rounded	St. Deviation	Median
Human Health	36%	40%	19%	33%
Ecosystem Quality	43%	40%	20%	33%
Resources	21%	20%	14%	23%

Source: PRé (Product Ecology Consultants), 2000

A weighting triangle can then be used to represent the results graphically (Hofstetter *et al*, 1999). Any relative weighting can be shown in the triangle. For each weighting set, the triangle then shows graphically which alternatives score best.

The results are shown in the figure below. Each point represents a combination of weights from a single panel member. The cross in the middle represents ten panel members, who gave equal weights. The dot represents the average rounded weighting.

Figure 6: Weighting of the damage categories by the panel



Source: PRé (Product Ecology Consultants), 2000

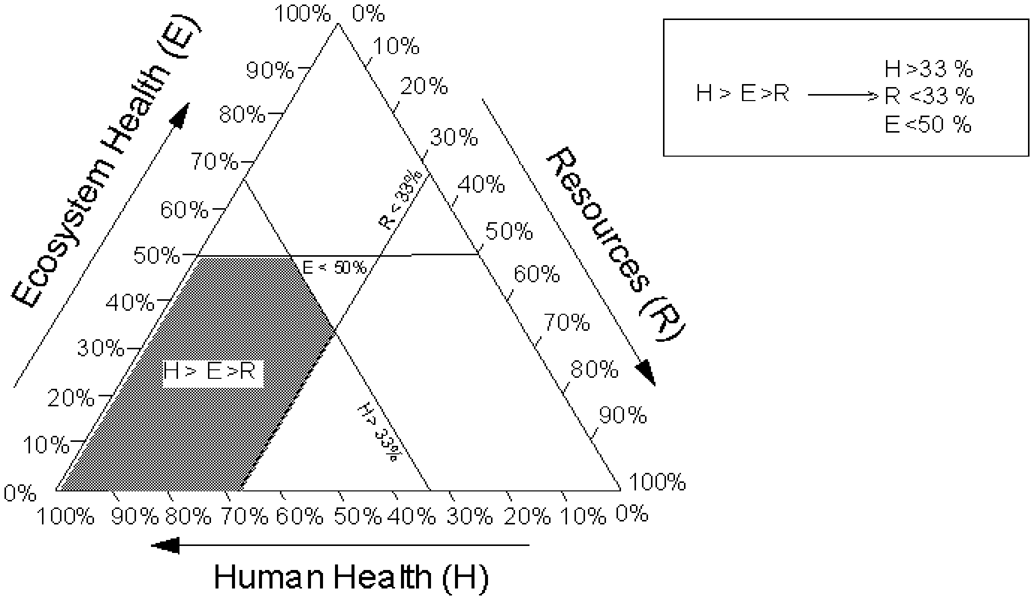
Next to weighting, the panel members were asked to rank the three damage categories according to their perceived importance. The triangle concept can be used to demonstrate the result of the ranking performed by the respondents.

If a panel member considers Human Health (H) to be more important than Ecosystem Quality (E) and that Ecosystem Quality is more important than Resources (R), this is interpreted as follows:

1. Human Health (H) has a weight that is higher than 33%. Otherwise E or R would by definition get the highest factor.
2. Resources has a weight that is lower than 33%. Otherwise it would be higher than either H or E.
3. Ecosystem Quality (E) has a weight lower than 50%. Otherwise it would be higher than H if the weight of R is zero.

This reasoning can be shown graphically in the triangle as the grey area in the figure below, representing the preferences of a single respondent.

Figure 7: Graphical representation of a ranking result



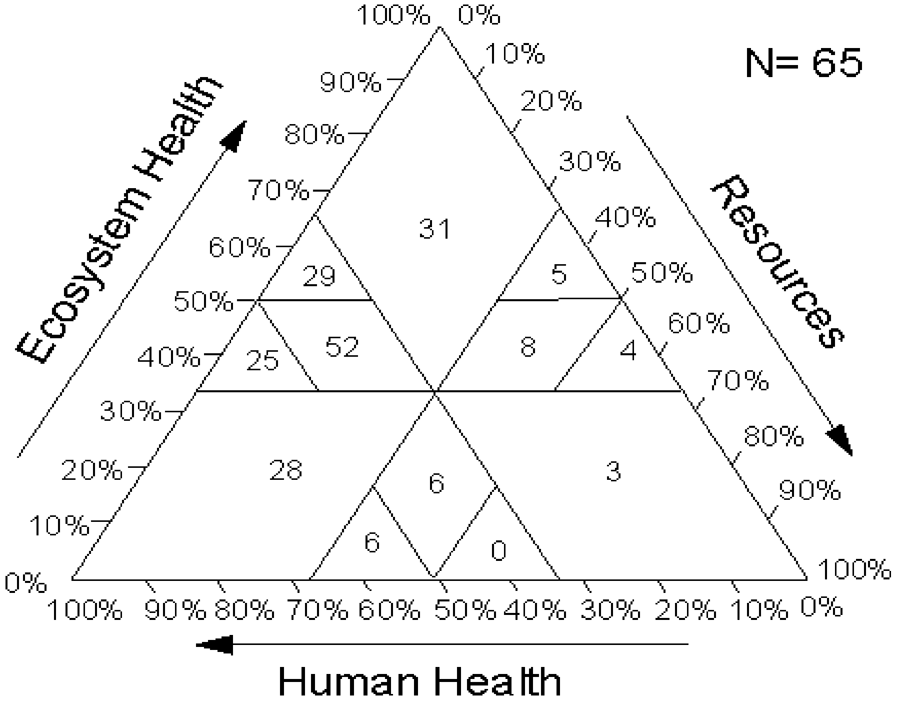
Source: PRé (Product Ecology Consultants), 2000

A respondent that ranks the damage categories as Human Health first, followed by Ecosystem Quality, followed by Resources, should have a weighting set that fits in the dark grey area.

With this reasoning the areas for all combinations that have been answered by all respondents can be plotted. In the following figure all areas for all respondents are combined in one triangle with the frequency by which an area or part of an area was chosen by the respondents.

This figure must however be interpreted with care, as the frequency fields overlap. In spite of this, the picture shows quite clearly that in the ranking Resources receive a relatively low weight, while Ecosystem Quality and Human Health receive approximately equal weights. This is in accordance with the results of the weighting procedure.

Figure 8: Overview of the ranking results of all respondents



Source: PRé (Product Ecology Consultants), 2000

This graphical representation of the weightings of indicators may be a useful way to make the weights more transparent. However, there are some limitations to the number of indicators for which weights can realistically be elicited and shown in this way.

5.5.4. Applying decision rules to form a composite

In many cases when the different (transformed) performance indicators are combined into a composite indicator, a set of decision rules are applied as to how the indicators should be combined. An example of this is in the construction of the scorecard for local authorities in the Comprehensive Performance Assessment. One of the rules for instance states that local authorities must score at least 3 (2 stars) on education, social services combined star rating, and financial standing to achieve a category of excellent overall. In the star ratings in health care a similar 6-step process is implemented using an algorithm to determine the ultimate star rating.

The primary reason for applying such decision rules is to try to ensure that certain minimum requirements must be met before managers become eligible for any further reward (for example, the requirement that the CHI clinical governance review is satisfactory). This implies on the part of the regulator an interest in certain minimum standards, or reducing disparities. From the managerial perspective, it can be modelled as a sequential decision problem (or lexicographic choice).

The manager therefore faces a set of standards or thresholds that are set by the regulator alongside a reward schedule which dictates the level of effort which will be expended to try to attain the standard and proceed to the next step in the set of decision rules. The manager's utility function will include the perceived reward or penalty and the perceived effort required to attain the standard.

Thus, using backward induction, the manager first calculates the maximum expected utility based on optimal actions supposing the threshold has been passed. This is compared with the utility when *not* meeting the threshold. Only if the former is greater than the latter, will the manager then go on to optimize the second stage behaviour.

The simple composite indicator (whether discrete or continuous) assumes that the marginal benefit of an extra unit of performance on any dimension is independent of performance in any other dimension. However, the reward schedule may be designed such that the portfolio of outcomes affects rewards. In its simplest form, this might require that the crude composite is amended depending on how many scores meet some minimum threshold.

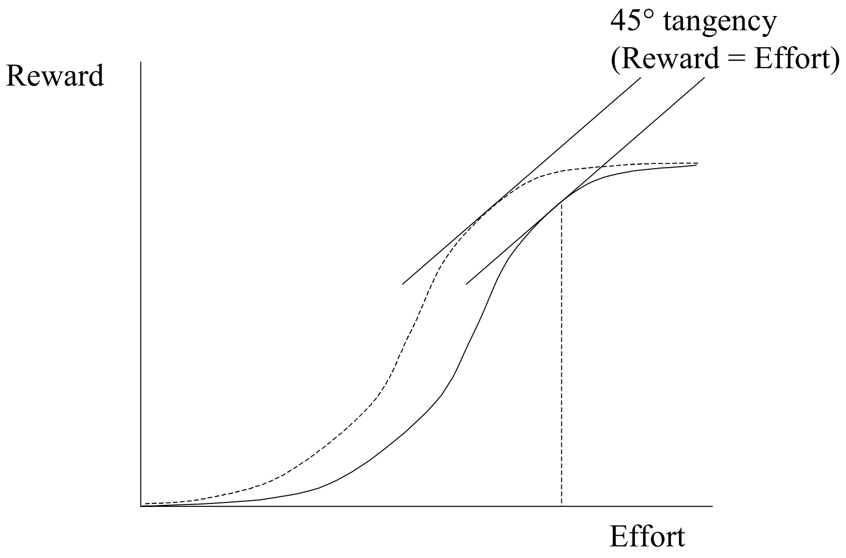
Suppose for example that the rule is that the reward is reduced if less than k performance measures satisfy some basic standards. Then the manager should calculate expected utility assuming optimal behaviour (a) satisfying the k-standard constraint and (b) removing the k-standard constraint. If (b) is greater, then the organisation will not seek to meet the standard.

Introducing uncertainty into the above may alter the incentives. If managers are risk averse then the rewards they require will have to be skewed accordingly (towards indicators with greater risk). Also, risk averse managers will change behaviour compared to the risk neutral case. In general:

- Risk aversion will reduce the probability that managers will 'go for' a threshold, but
- If they do want to meet the threshold, they will seek to overshoot to 'be on the safe side'
- They will balance the increased effort of overshooting against the reduced probability of failure
- These effects will be exaggerated by increased uncertainty.

The following figure illustrates the issue. The probability of meeting the threshold increases with effort. The dotted curve shows the risk neutral situation, and the optimum (marginal reward equals marginal effort) is at the point of 45^0 tangency. The solid curve is the risk averse situation (the curve is lowered by the risk premium). Note that effort is increased under risk aversion. But because more effort is needed there is an increased probability that the optimum yields less utility than the 'zero effort' case and the agent therefore decides not to participate.

Figure 9: Expected reward under risk aversion



Therefore, the application of decision rules to form a composite may have important incentive effects influencing how managers will respond to the minimum service standards or thresholds, depending on the reward schedule, required effort and their degree of risk aversion.

5.6. Adjusting for environmental or uncontrollable factors

Some units must operate in more adverse environmental circumstances which may make the attainment of performance outcomes more difficult for them. Thus for a given level of expenditure, the production possibility frontiers for these systems will lie inside those with more favourable environmental conditions (Smith, 2002). It is therefore argued that adjustments might be made to take into account these exogenous environmental conditions when measuring their performance.

There may be many causes of variation and exogenous influences on performance (Jacobs & Dawson, 2002; Gravelle & Smith, 2002).

Some of these may include:

1. differences in health status of the population (for example age and gender mix, co-morbidities, case-mix, and so on)
2. the external environment (for example geography),
3. differences in resources used (or an inappropriate mix of resources),
4. differences in the quality of services provided,
5. variations in institutional arrangements (for example specific hospital features),
6. different priorities regarding objectives,
7. different accounting treatments (and data conventions),
8. data errors,

9. random fluctuations (some measures may be more vulnerable to fluctuations beyond the control of the system), and
10. differences in the effectiveness of the system in achieving the chosen objectives – the key issue of interest.

Thus composite indicators of the NHS agencies may not just be measuring NHS performance but also population characteristics such as unemployment, education, housing conditions and so on which influence health status and the use of health services.

It may not always be possible or policy relevant to correct for all these sources of variation under all circumstances. Of key importance here is the extent to which environmental influences (such as 1 and 2) are taken into account, by adjusting for these differences in environment as they would impact on the capability of units to deliver performance equally.

The debate around whether exogenous environmental factors should be taken into account really boils down to delineating the boundaries of the health care system under investigation. If the composite is focused on performance of the whole health care 'system' (all activities and actions by government and others) then there is less argument for standardising confounding factors. There is an argument for not treating these and other influences on the composite as exogenous, if this broader perspective is taken which encompasses a broad view of health and health care. This wider view argues that though NHS agencies may be unable to influence certain factors such as unemployment, these factors can be influenced by other parts of government (Appleby & Mulligan, 2000).

However, if the interest is in a more narrow definition of health care agencies, then it may be very important to adjust for all exogenous contextual variables beyond the control of the unit. If the composite is focused on performance of certain NHS units alone, then exogenous factors beyond the direct influence of these agencies should be controlled for or standardised.

It could be argued that if the English health care funding formula used to distribute funds to health authorities then or Primary Care Trusts now, for example, is designed to enable all health delivery units under investigation to produce equal levels of performance, given their different environmental circumstances, then there may be no need to adjust for exogenous circumstances in the construction of the composite (Smith *et al*, 2001). These formulae take into account population characteristics and it can be argued that an indirect form of standardisation is therefore carried out via the funding system. This assumes that purchasers are adequately compensated for their differences in health needs.

Thus if the funding formula is correctly designed to permit equal performance, and all relevant aspects of performance could be captured by the composite, then there would be no need to control for exogenous factors, or indeed include cost in any efficiency models. Efficiency need only then be examined as the difference between observed outputs or outcomes of different units.

There are generally two different levels at which adjustment for uncontrollable influences on performance can take place. The first is at the final stages after having constructed the composite indicator, the second, which is not mutually exclusive to the first, is to (also) adjust the individual performance indicators for differences in the underlying population risk.

In the first case where adjustment is made at the level of the composite, there may be technical difficulties in trying to incorporate exogenous influences on performance. There are in the productivity literature generally two approaches to modelling exogenous influences on a set of performance scores. The first is to incorporate them in a one-step model where they are included as an input in the production process. The second is a two-step model where the model is first set up to exclude exogenous factors and then in a second step the performance scores are explained by exogenous influences. While the two-step approach may be more transparent and practical, it is often contentious which factors are considered truly exogenous and should be left out of the first step. In addition, the variables in the first and second stages are likely to be highly correlated, leading to biased results (Simar and Wilson, 2002). Furthermore, scores from the first step will be sensitive to whether or not some exogenous factors have been included (Fried *et al*, 1999).

The essential point is that there is no generally accepted method for taking into account environmental variables at the level of the composite scores, or for testing whether an environmental variable has a significant influence on the production process and the resultant performance of the unit.

In the second case, individual indicators are often adjusted for differences in the health status of the population. Risk adjustments are made to the data to include age, sex, type of admission, length of stay and co-morbidity profiles of the relevant population. However there are also technical difficulties with this approach. Alternative methods of risk-adjustment usually give rise to different results and may lead to large variations in performance assessment (Iezzoni *et al*, 1995; Iezzoni *et al*, 1996). Furthermore, as technology and clinical competence change, associated risk adjustment schemes must change. In some specialties for some disease groups outcome data are inappropriate (such as psychiatry, rheumatology, dermatology or ophthalmology) (Davies and Crombie, 1995). Risk adjustments to indicators measuring performance in these areas may therefore need to take a different form to other areas of health care.

One way of examining units facing a wide variation in environmental circumstances, is to present the results in the form of clusters of comparable units rather than as a single index or league table. This may be particularly appropriate where an index spans many different types of organisations or countries. For example, the environmental sustainability index is also reported in terms of 5 clusters of countries with similar profiles and cluster analysis also revealed interesting patterns in the average values of their scores, values of the scores on components of the composite and other factors thought to be important (such as degree of democracy) (World Economic Forum, 2002). This analysis allowed the authors to suggest reasons for variations in performance between groups and made the policy implications much more transparent than a comparison of the index across all countries would allow.

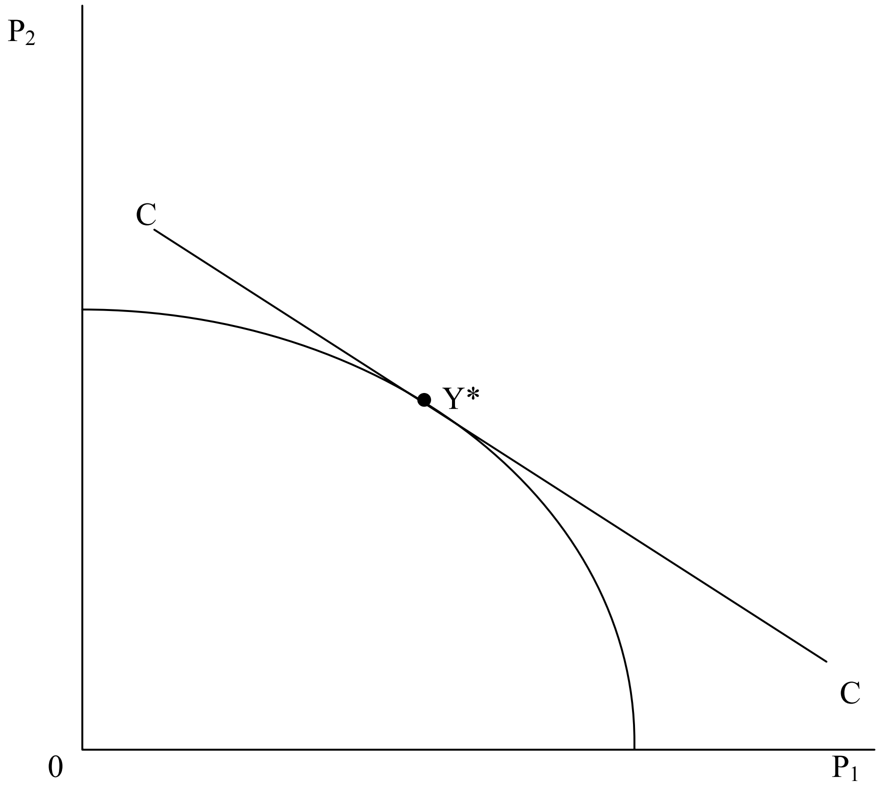
5.7. Examining variations in efficiency

As a final step in the construction of a composite indicator, regulators may be interested in exploring the efficiency with which organisations use resources in relation to achieving the performance measured on the composite. This leads to the examination of performance in relation to some measure of resource use, usually cost (Smith, 2002). This allows economists to examine the ratio between performance (outputs or outcomes) and resource use or costs devoted to the attainment of the performance (inputs) which is typically a measure of efficiency.

There are two broad approaches for analysing efficiency, namely stochastic frontier analysis (SFA) and data envelopment analysis (DEA). In these methods, performance of a system is modelled as a function of resource use (expenditure) and any relevant environmental factors and efficiency is inferred from an empirical production frontier. The degree to which a particular unit exceeds or falls short of the predicted performance based on the production frontier, determines its relative efficiency. Technical choices can dramatically affect the relative efficiency of individual units and the two techniques can generate very different results (Jacobs, 2001).

The approach of examining efficiency relative to a production possibility frontier is illustrated in the following diagram for a composite performance measure.

Figure 10: The production possibility frontier with two outputs



Assuming two outputs (and a single organization), with a fixed budget constraint, an unconstrained organisation would choose a point on the frontier in accordance with its objective function, which incorporates the relative weighting of outputs 1 and 2.

The use of a simple linear composite indicator can be illustrated by straight lines with slopes as in CC. In this example, an organisation that wishes to maximize its composite score will choose point Y^* on the frontier. An organisation that secures a composite score less than CC displays either technical inefficiency (it lies within the frontier), or allocative inefficiency (it lies on the frontier, but not at Y^*), or some mix of technical and allocative inefficiency.

This analysis suggests that the budget given to the organisation should be informed by the performance measurement regime. In particular, if organisations are to be ranked against their composite scores, then they should be given budgets that in some sense give them equal opportunities to secure equal composite scores. One example of such a budgetary regime would be to give every organization a budget that allows them just to achieve composite score CC if they are technically and allocatively efficient.

In principle, setting such 'fair' budgets requires full knowledge of each organization's multi-output production function. In general, production functions vary between organizations depending on the environmental difficulties that confront them. In practice, budgets in the UK public services are often set according to 'average' expenditure levels for organizations confronted by similar environmental circumstances (Smith *et al*, 2001). This probably achieves some element of horizontal equity (organizations in similar circumstances are treated equally). However, because this method of setting budgets merely replicates current spending patterns, it does not necessarily secure the vertical equity needed to be able to rank the performance of organizations in different environmental circumstances on a consistent basis.

For example, if currently organisations (such as Primary Care Trusts) in adverse environments are generally scoring poorly on a composite measure relatively to their less disadvantaged counterparts, this may be because they are consistently less efficient. However, it may also be because they are not currently funded well enough to secure higher scores. If this is the case, the funding formula needs to skew resources further towards organizations in adverse circumstances in order to offer them a level playing field (Hauck *et al*, 2002; Smith, 2003). Integrating performance criteria with funding formulae may therefore require quite radical revisions to the methodology for setting 'fair' budgets.

5.8. Sensitivity analysis of the construction of the composite indicator

As seen from each of the preceding steps in constructing a composite indicator, there are a variety of difficulties that can arise with respect to selecting, transforming, weighting and aggregating variables into a composite. The outcomes and rankings of individual units on the composite may largely depend on the decisions taken at each of the preceding steps. As such, an important consideration is the use of a sensitivity analysis to explore the robustness of rankings to the inclusion and exclusion of certain variables, changes in the weighting system, using different transformation methods and setting different decision rules to construct the composite (Freudenberg, 2003).

While most analysts would probably agree that sensitivity analysis is considered good practice, in reality this is seldom exercised.

Sensitivity analysis can be done by examining the rank and score correlation coefficients of units across different composite indicators using, for instance, different transformation techniques or different choices of weights. As an example, an analysis of the impact of using sets of weights derived from different sources in estimating an indicator of technology achievement, showed that in many cases, the rankings overlapped (Joint Research Centre, 2002). Further analysis was able to pinpoint the specific weights to which the results were more sensitive.

In addition, a useful addition to the exercise would be to construct confidence intervals around the composite indicator since large standard errors around the composite would produce wide confidence intervals and a greater imprecision around the estimates. If the confidence intervals are overlapping across the entire series of units, then a great deal of caution should be exercised in attributing differences in apparent rankings of units to true differences in performance since these may be entirely spurious. In this case, using the composite league table as a regulatory tool and basing resource allocation and other important policy decisions on the results of the composite outcomes may be premature. Confidence intervals enable the important distinction to be made between differences in performance which are due to sampling error and natural variation, and true differences in performance for which managers may be held responsible. Being able to separate out this random variation, the result of measurement error or natural variability, will give much greater precision to the final composite.

6. Empirical analysis

This section explores some of the technical issues involved in the construction of a composite indicator using data from the Star Ratings system for acute hospital trusts in England. As mentioned, there is now three years of data available and whilst the methodology has remained relatively constant there have been some important changes to the underlying indicators chosen, the domains covered, the role of the CHI clinical governance review and the way in which the indicators have been combined to form the star rating. The data covers the years 2000/01 to 2002/03. This report uses the data published in 2002 for the year 2001/02.

In the empirical analysis, there are two main components to the work. The first part is exploratory and examines the underlying nature of the data, including the distributions of the underlying indicators, how they have been transformed, the correlations between the indicators and a factor analysis. The second part of the empirical work then uses the dataset to construct a new composite index through a simulation exercise. Each of the steps of constructing a new composite are simulated and the robustness of the rankings of individual hospital trusts are examined.

6.1. Data analysis

6.1.1. The underlying performance indicators

The data analysis focuses on the 2001/02 data which contains 38 performance indicators. Table 35 in the appendix gives a full list of all the variables and their definitions which are grouped into the key targets, the three domains of clinical, patient, and capacity and capability focus and the CHI review. The three domains of performance indicators make up the 'balanced scorecard'.

Table 36 in the appendix gives the descriptive statistics for the raw performance indicators prior to being transformed. These include the number of observations, the mean, median and standard deviation as well as measures of skewness and kurtosis which give an indication of the type of distribution of the variable. Skewness is a measure of the lack of symmetry of a distribution. If the coefficient of skewness is zero, the distribution is symmetric. If the coefficient is negative, the median is usually greater than the mean and the distribution is skewed left. If the coefficient is positive, the median is usually less than the mean and the distribution is skewed right. Kurtosis is a measure of peakedness of a distribution. The smaller the coefficient of kurtosis, the flatter the distribution. The normal distribution has a coefficient of kurtosis of 3 and provides a convenient benchmark. The test for normality based on D'Agostina *et al* (1990), runs separate tests for normality based on skewness and kurtosis and then combines the two tests into an overall Chi-squared test statistic.

Several of the variables have a Prob > Chi-squared = 0.000 which suggests they are significantly skewed. This is particularly the case with all the key targets. Quite a large number of the indicators within the balanced scorecard however appear to have approximately normal distributions.

It is interesting to note that improved working lives (taking a value of 1 only) and cleanliness (taking values of 3 or 4 only) have very little variation. As mentioned these are threshold variables, typical of key targets, which measure service standards. The inability to run normality tests on these variables is indicative of the lack of variation on these indicators.

Of the clinical indicators, general readmission rates (readmisnpc) has a lower standard deviation than the other readmission rates (for hips, stroke and children). This is because it is based on large numbers of events and hence has smaller standard errors producing narrower confidence intervals.

It is important to note that two of the indicators (deaths from heart bypass and heart operation waits) have very small numbers of observations. This highlights the issue of the level of data aggregation, where a large number of hospitals do not perform heart surgery and cannot be measured on the included indicator.

It is also interesting to note that the type of indicator and its distribution need not be related to the type of performance which is being measured. For instance, there are a large number of different types of waiting times measures included. However, some are threshold variables (the long wait key targets) whilst others are continuous and seem to measure differential performance across hospital trusts (such as A&E waits, outpatient waits and 6 month inpatient waits). Thus the actual performance which is being measured need not dictate the type of indicator which can be used. If an indicator is deemed important to include (such as cleanliness) but is displaying little variation, it need not be excluded, but could be measured in a different way, either on a different scale or using a different threshold, if it is considered important to pick up differential performance.

6.1.2. The transformed variables

Table 37 in the appendix shows the thresholds which were applied to transform the raw continuous variables into categorical variables in the star ratings.

Both quantitative and qualitative categorical scales are used, for instance 'achieved', under achieved' and 'significantly under achieved' and also a three or five point scale. Since the clinical indicators are published with confidence intervals, they are transformed into a 1, 3 or 5 rating depending on whether they are below, on or above the English average on the confidence interval.

As mentioned the choice of thresholds to transform the variables varies across the different indicators. For instance, there is not a universal rule of using a uniform percentile banding. The thresholds for deciding the cut-offs for each of these categories varies for each variable, to ensure that the resulting transformed categorical variable has an approximately normal distribution, even if the underlying raw performance indicator or key target did not. As such, the scales can be considered highly subjective since the choice of thresholds can be selected arbitrarily.

The following Table 38 in the appendix gives the descriptive statistics for the transformed variables which are either on a three or five point scale. It is evident that for some indicators such as improved working lives and cleanliness there is no variation between Trusts, making the indicator essentially useless for discerning differential performance. As mentioned, this is because of the threshold chosen to measure the performance.

The following figures give examples of how some of the raw data (the variables on the left) have been transformed into categorical variables (the variables on the right) and how the distributions of the variables are dramatically changed by applying these different thresholds.

Thus the bottom two figures (breast cancer waits and total inpatient waits) have significantly skewed distributions to the left and right respectively. The top variable delayed discharges has a much more symmetrical distribution. However, because of the thresholds chosen, each of the distributions on the right turn out much more 'normal'. In particular the outlier extreme data in the bottom two figures is included in the same categories as those on the edge of the distribution. Thus the incentives for improving on (say) extremely poor performance are minimal.

Figure 11: Examples of transformations of variables and threshold choices

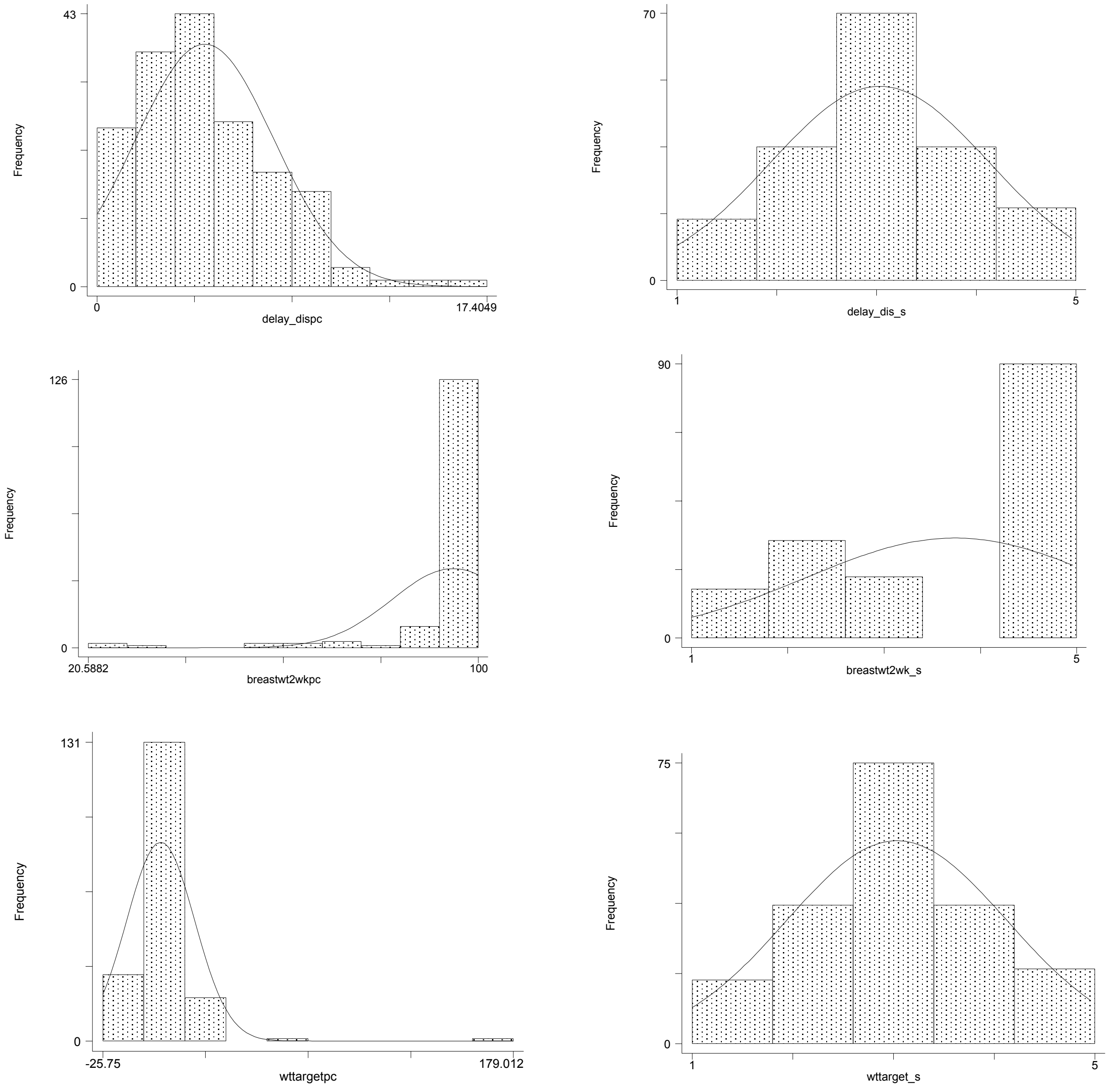


Table 39 in the appendix gives the frequency and percent distributions for the transformed variables. These essentially match the above transformed figures on the right hand side.

6.1.3. Correlations between indicators

The next step in the data analysis was to explore the collinearity between the different indicators. For this, the raw (untransformed) data was used.

It is expected that a number of the indicators will be highly correlated, since as mentioned, ‘good’ organisations will tend to score well across most indicators (that is, cov(x_i,x_j)>0 for most (*i,j*) from equation (13)).

The correlation matrix for all 38 performance indicators was produced but is too complex to include in the report in its entirety. Hence the following sets of tables show sub-sets of correlations for various indicators.

As mentioned, three of the indicators had very low numbers of observations which significantly reduced the numbers of observations in the correlations in which these variables were included. Hence the correlations were run with and without these three variables (deaths from heart bypass, heart operation waits and the CHI review).

The following table shows the correlations for all waiting times variables including heart operation waits. There are significant correlations between the inpatient and outpatient waiting times and between A&E and outpatient waits, which may allude to issues of bed capacity. The two types of cancer wait measures are also highly correlated. Correlations above ±0.4 are highlighted.

Table 18: Correlations between waiting times, n=26

	inpwt18mn	outwt26wk	a_e12hrwt	cancerwt2wkpc	wait6pc	outwt13wkpc	a_e4hrwtpc	wt_heart	breastwt2wkpc
inpwt18mn	1.000								
outwt26wk	-0.059	1.000							
a_e12hrwt	0.393	-0.057	1.000						
cancerwt2wkpc	-0.079	0.098	-0.301	1.000					
wait6pc	-0.307	-0.030	0.143	-0.123	1.000				
outwt13wkpc	-0.370	-0.079	-0.088	0.345	**0.451**	1.000			
a_e4hrwtpc	-0.079	**0.714**	-0.075	0.188	-0.135	-0.144	1.000		
wt_heart	-0.055	-0.043	-0.054	0.188	-0.068	0.003	-0.058	1.000	
breastwt2wkpc	0.028	0.106	-0.034	**0.531**	-0.002	0.163	0.151	0.108	1.000

The following table shows the same variables but excluding heart operation waits which has very few observations. The correlations between these same variables now drop substantially.

Table 19: Correlations between waiting times, n=145

	inpwt18mn	outwt26wk	a_e12hrwt	cancerwt2wkpc	wait6pc	outwt13wkpc	a_e4hrwtpc	breastwt2wkpc
inpwt18mn	1.000							
outwt26wk	0.010	1.000						
a_e12hrwt	0.002	0.317	1.000					
cancerwt2wkpc	-0.023	-0.031	-0.192	1.000				
wait6pc	-0.049	-0.134	-0.135	0.227	1.000			
outwt13wkpc	-0.085	0.051	-0.055	0.276	0.310	1.000		
a_e4hrwtpc	0.003	-0.151	-0.178	-0.032	0.190	0.267	1.000	
breastwt2wkpc	-0.008	0.040	0.013	0.373	0.085	0.040	-0.026	1.000

The following table shows the correlations between the clinical indicators and includes deaths from heart bypass and hence has fewer observations. There appear to be significant associations between clinical negligence and death rates, as well as between death rates and readmission rates.

Table 20: Correlations between clinical indicators, n=21

	cnst	d_esurgstd	d_heartpl	readmisnpc	readm_child	readm_hip	readm_stroke	dis_hippc	dis_strokepc	delay_dispc
cnst	1.000									
d_esurgstd	**-0.506**	1.000								
d_heartpl	**-0.645**	**0.422**	1.000							
Readmisnpc	-0.271	0.185	0.296	1.000						
readm_child	-0.126	**0.408**	0.041	0.097	1.000					
readm_hip	0.066	0.153	0.185	0.268	-0.298	1.000				
readm_stroke	-0.098	-0.131	0.347	-0.111	-0.039	0.200	1.000			
dis_hippc	-0.010	-0.072	-0.183	0.000	-0.168	0.097	0.098	1.000		
dis_strokepc	0.236	0.080	-0.332	-0.265	-0.111	0.188	0.097	0.313	1.000	
delay_dispc	0.016	0.041	-0.073	-0.102	0.098	0.209	-0.179	-0.382	0.063	1.000

Once again, dropping the clinical indicator for deaths from bypass operations provides a bigger sample size and many of the correlations drop substantially again. There is a stronger correlation again between readmission rates for hip fractures and strokes which may relate to issues of case-mix.

Table 21: Correlations between clinical indicators, n=125

	cnst	d_esurgstd	readmisnpc	readm_child	readm_hip	readm_stroke	dis_hippc	dis_strokepc	delay_dispc
cnst	1.000								
d_esurgstd	-0.086	1.000							
readmisnpc	-0.028	0.015	1.000						
readm_child	-0.012	0.100	0.288	1.000					
readm_hip	-0.003	0.064	0.381	-0.025	1.000				
readm_stroke	-0.050	-0.042	0.187	0.001	**0.490**	1.000			
dis_hippc	-0.029	0.020	0.089	0.047	0.127	0.097	1.000		
dis_strokepc	-0.021	-0.170	0.265	0.126	0.196	0.138	0.202	1.000	
delay_dispc	0.017	0.062	-0.271	-0.015	-0.031	0.038	0.039	-0.150	1.000

Excluding deaths from heart bypass and heart operation waits, the following table shows the correlations for the waiting times and readmission rates. Again, significant correlations are observed between the different indicators of cancer waits and readmission rates for hip fractures and strokes.

Table 22: Correlations between waiting times and readmission rates, n=127

	cancerwt2wkpc	wait6pc	outwt13wkpc	a_e4hrwtpc	breastwt2wkpc	readmisnpc	readm_child	readm_hip	readm_stroke
cancerwt2wkpc	1.000								
wait6pc	0.171	1.000							
outwt13wkpc	0.235	0.310	1.000						
a_e4hrwtpc	0.003	0.225	0.280	1.000					
breastwt2wkpc	**0.426**	0.041	0.076	-0.002	1.000				
readmisnpc	-0.034	0.147	-0.059	0.139	-0.092	1.000			
readm_child	-0.069	0.040	-0.092	-0.061	-0.031	0.328	1.000		
readm_hip	0.098	0.171	0.097	-0.039	-0.065	0.316	0.003	1.000	
readm_stroke	-0.136	0.052	-0.027	-0.056	-0.002	0.198	0.004	**0.406**	1.000

The following table shows the correlations between the various measures of patient satisfaction and a number of staff satisfaction variables. All the inpatient satisfaction variables are highly correlated suggesting they pick up similar measures of satisfaction with the inpatient experience.

Table 23: Correlations between patient satisfaction and staff variables, n=139

	inp_survey_coord	inp_survey_env	inp_survey_inf	inp_survey_phys	inp_survey_acc	inp_survey_resp	staff_survey	jundocpc	sick_rate
inp_survey_coord	1.000								
inp_survey_env	**0.653**	1.000							
inp_survey_inf	**0.710**	**0.559**	1.000						
inp_survey_phys	**0.763**	**0.564**	**0.816**	1.000					
inp_survey_acc	**0.783**	**0.565**	**0.686**	**0.687**	1.000				
inp_survey_resp	**0.663**	**0.611**	**0.601**	**0.656**	**0.587**	1.000			
staff_survey	0.007	-0.066	0.091	0.013	-0.112	-0.046	1.000		
jundocpc	0.136	0.151	-0.053	0.050	0.167	0.128	-0.187	1.000	
sick_rate	-0.008	0.237	-0.053	-0.002	0.017	0.129	-0.158	0.296	1.000

The following table shows the correlations between other variables not previously included in the above tables, including the CHI review. The inclusion of the latter variable dramatically reduces the sample size again. This indicator is, unsurprisingly, highly correlated with the star ratings, since it plays a significant role in the algorithm or decision rules used to generate the star ratings. Different measures of cancelled operations are also highly correlated.

Table 24: Correlations between all other variables, n=84

	pi_stars	cancelopspc	cleanliness	finbalpc	cancelop1mnpc	delay_dispc	dqi_pc	info_gov	chi_review
pi_stars	1.000								
cancelopspc	-0.371	1.000							
cleanliness	0.120	0.106	1.000						
finbalpc	0.187	-0.011	-0.117	1.000					
cancelop1mnpc	-0.278	**0.805**	0.114	0.028	1.000				
delay_dispc	-0.065	-0.008	0.071	-0.110	-0.011	1.000			
dqi_pc	0.063	0.148	-0.077	0.191	0.087	0.013	1.000		
info_gov	0.184	-0.093	0.281	-0.138	-0.172	0.069	0.198	1.000	
chi_review	**0.704**	-0.198	0.187	0.020	-0.095	0.044	0.004	0.255	1.000

In the above tables, the results for the correlations with larger sample sizes seem to provide more robust results on the likely associations between different types of performance measures. In general, these seem to pick up stronger correlations between readmission rates for hip fractures and strokes, different indicators of cancer waits, various measures of patient satisfaction, different measures of cancelled operations, and the star ratings and CHI review.

6.1.4. *Factor analysis*

In this section, one of the main methods for exploring the relationship between different indicators is used, namely factor analysis.

Factor analysis is essentially a data reduction method with the principal idea being that one can describe a set of p variables $X_1, X_2, ..., X_p$ in terms of a smaller number of *m* factors. Thus each of the variables takes the form $X_i = a_i F + e_i$ where X_i is the i^{th} standardised score with a mean of zero and a standard deviation of one, a_i is a constant, F is a 'factor' value with a mean of zero and a standard deviation of one and e_i is the part of X_i that is specific to the i^{th} score only. As a consequence of these assumptions, a constant ratio between the rows of a correlation matrix follow and there is a plausible model for the data. A general form of this model is as follows:

$$\begin{aligned} X_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + \ldots + \alpha_{1m}F_m + e_1 \\ X_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + \ldots + \alpha_{2m}F_m + e_2 \\ &\vdots \\ X_p &= \alpha_{p1}F_1 + \alpha_{p2}F_2 + \ldots + \alpha_{pm}F_m + e_p \end{aligned} \qquad (17)$$

where X_i is a variable with a mean of zero and a standard deviation of one, $\alpha_{i1}, \alpha_{i2}, \ldots \alpha_{im}$ are the factor loadings related to the variable X_i, $F_1, F_2, \ldots, F_m$ are *m* uncorrelated common factors, each with a mean of zero and a standard deviation of one and e_i is the specific factor related only to the variable X_i, which has zero mean and is uncorrelated with any other common factor and the specific factors (Joint Research Centre, 2002).

Rotated factor analysis simply produces results which can more readily be interpreted. If factor loadings or correlations could be produced on a plot, with each variable represented as a point, the axes of this plot could be rotated in any direction without changing the *relative* locations of the points to each other. However, the actual coordinates of the points, that is, the factor loadings would of course change. There are various rotational strategies that have been proposed. The goal of all of these strategies is to obtain a clear pattern of loadings, that is, factors that are somehow clearly marked by high loadings for some variables and low loadings for others. Thus a rotation is sought that maximizes the variance on the new axes and thus produces a pattern of loadings on each factor that is as diverse as possible, lending itself to easier interpretation.

Since factor analysis is based on a correlation or covariance matrix, it assumes the observed indicators are measured continuously, are distributed normally, or at least symmetrically, and that the associations among indicators are linear. That said, exploratory factor analysis is often used as a data reduction technique with ordered categorical indicators and dichotomous variables. While this report has shown that many of the variables are highly skewed and not all are continuous (although the untransformed non-categorical variables are used), this exercise is strictly exploratory to examine the associations between indicators, and hence all variables have been included, irrespective of their underlying distributions.

The following table shows the rotated factor loadings for all the performance indicators excluding deaths from heart bypass and heart operation waits, since these reduced the sample size too much.

The uniqueness factor indicates the proportion of a variable's variance that is not shared with a factor structure and is therefore a measure of its statistical 'uniqueness'.

Table 25: Rotated factor loadings for all variables

Variable	1	2	3	4	5	6	Uniqueness
inpwt18mn	-0.051	**0.982**	0.041	-0.020	0.032	-0.009	0.002
wt15mn	0.005	**0.993**	0.020	-0.034	0.001	-0.022	-0.004
outwt26wk	-0.066	**0.693**	0.064	-0.149	-0.049	-0.055	0.204
a_e12hrwt	-0.051	**0.745**	0.064	-0.015	0.047	-0.011	0.200
cancelopspc	-0.119	0.073	**0.893**	-0.017	-0.014	-0.037	0.111
cancerwt2wkpc	0.137	-0.102	0.036	-0.058	0.091	-0.020	0.173
cleanliness	-0.057	-0.108	0.202	-0.171	0.065	0.102	0.343
finbalpc	-0.087	0.331	-0.006	0.028	0.008	0.001	0.249
cnst	-0.027	-0.300	0.191	-0.058	-0.072	0.313	0.284
d_esurgstd	-0.147	-0.197	0.144	0.193	0.307	0.045	0.450
readmisnpc	0.076	-0.083	-0.029	0.256	0.146	0.036	0.162
readm_child	0.104	0.104	0.025	-0.032	**0.724**	0.034	0.303
readm_hip	-0.035	-0.190	0.113	**0.738**	0.018	0.009	0.189
readm_stroke	-0.019	-0.015	-0.181	**0.850**	-0.026	-0.056	0.158
dis_hippc	-0.097	-0.006	0.038	0.006	0.130	0.061	0.336
dis_strokepc	0.088	-0.100	-0.036	0.134	-0.034	0.163	0.279
wait6pc	0.068	-0.312	0.102	0.080	0.058	-0.050	0.141
wttargetpc	0.178	0.255	0.187	-0.065	0.135	-0.072	0.243
outwt13wkpc	0.287	0.214	-0.175	-0.004	-0.251	0.071	0.229
a_e4hrwtpc	0.284	-0.240	-0.272	-0.041	-0.072	0.104	0.129
cancelop1mnpc	-0.117	0.047	**0.911**	-0.067	0.039	0.009	0.125
breastwt2wkpc	0.082	0.057	0.050	0.037	0.001	0.101	0.246
delay_dispc	-0.207	-0.077	-0.087	-0.099	-0.053	0.169	0.240
inp_survey_coord	**0.716**	-0.052	-0.203	-0.137	0.079	0.123	0.172
inp_survey_env	**0.750**	0.078	-0.024	0.029	0.266	0.096	0.078
inp_survey_inf	**0.786**	-0.082	-0.085	-0.080	0.022	-0.049	0.154
inp_survey_phys	**0.937**	0.020	0.015	0.009	-0.078	0.086	0.047
inp_survey_acc	**0.689**	-0.158	-0.207	0.064	0.049	-0.078	0.034
inp_survey_resp	**0.781**	-0.076	-0.083	0.037	-0.075	-0.040	0.163
dqi_pc	0.089	-0.071	0.117	0.037	-0.016	0.070	0.214
staff_survey	-0.105	-0.041	-0.272	0.127	0.061	0.158	0.168
jundocpc	0.233	-0.210	0.029	0.027	0.047	0.174	0.226
sick_rate	0.183	-0.097	0.067	0.130	0.146	-0.097	0.270
info_gov	0.112	-0.286	-0.106	-0.079	0.063	0.309	0.319
chi_review	0.107	-0.038	-0.034	-0.041	0.025	**0.848**	0.191

A clear set of distinct factors emerge which seem to accord with the previous correlation results. These factors can be broadly interpreted as:

- Inpatient satisfaction
- Waiting times
- Cancelled operations
- Readmission rates (for adults)
- Readmission rates (for children)
- The CHI review

A second set of rotated factor loadings (correlations) are shown in the following table, this time also excluding the CHI review which tends to reduce the sample size substantially since not all trusts had received a CHI review.

Table 26: Rotated factor loadings for all variables excluding CHI review

Variable	1	2	3	4	5	6	7	Uniqueness
inpwt18mn	-0.009	0.035	-0.033	-0.211	-0.015	0.016	0.061	0.743
wt15mn	-0.020	**0.881**	0.034	-0.024	-0.016	0.030	-0.053	0.192
outwt26wk	-0.050	**0.752**	0.058	-0.086	0.042	0.020	-0.057	0.334
a_e12hrwt	0.015	**0.506**	0.129	-0.036	-0.169	-0.138	0.378	0.428
cancelopspc	-0.162	0.049	**0.833**	-0.037	-0.054	-0.050	0.009	0.257
cancerwt2wkpc	0.027	-0.078	-0.008	-0.018	**0.699**	0.183	-0.048	0.428
cleanliness	0.070	-0.094	0.156	-0.135	0.005	0.067	0.032	0.510
finbalpc	0.030	0.184	0.018	-0.019	-0.043	-0.015	**-0.628**	0.544
cnst	0.090	-0.181	0.130	-0.110	-0.113	0.162	0.124	0.644
d_esurgstd	-0.252	-0.155	0.044	0.073	0.079	-0.007	-0.084	0.573
readmisnpc	0.129	-0.018	-0.001	0.272	0.009	-0.042	-0.051	0.416
readm_child	0.159	0.089	0.007	-0.045	0.165	-0.179	-0.051	0.589
readm_hip	-0.097	-0.161	0.017	**0.672**	0.028	0.190	0.128	0.383
readm_stroke	-0.032	-0.015	-0.129	**0.694**	-0.053	-0.134	-0.101	0.434
dis_hippc	-0.072	-0.014	0.119	0.092	-0.015	-0.110	-0.094	0.643
dis_strokepc	0.156	-0.064	-0.001	0.149	-0.247	0.024	-0.018	0.543
wait6pc	0.075	-0.198	0.005	0.046	0.140	0.114	-0.022	0.439
wttargetpc	0.107	0.186	0.221	-0.059	0.031	-0.144	0.069	0.501
outwt13wkpc	0.219	0.114	-0.207	0.053	0.191	**0.537**	-0.024	0.405
a_e4hrwtpc	0.434	-0.185	-0.218	-0.053	-0.045	0.128	-0.209	0.378
cancelop1mnpc	-0.105	0.060	**0.820**	-0.028	0.020	-0.008	-0.001	0.296
breastwt2wkpc	0.050	0.051	-0.065	-0.006	**0.630**	-0.123	0.081	0.529
delay_dispc	-0.260	-0.072	-0.061	0.045	-0.181	0.071	0.140	0.506
inp_survey_coord	**0.807**	-0.031	-0.102	-0.134	-0.030	0.122	-0.093	0.227
inp_survey_env	**0.723**	0.051	0.010	0.034	0.030	0.100	0.049	0.222
inp_survey_inf	**0.782**	-0.001	-0.092	-0.008	0.062	-0.136	0.134	0.235
inp_survey_phys	**0.824**	0.038	0.016	-0.069	0.016	-0.051	0.128	0.263
inp_survey_acc	**0.787**	-0.129	-0.178	0.025	0.070	0.030	-0.202	0.147
inp_survey_resp	**0.679**	0.011	-0.017	0.143	-0.116	0.069	-0.056	0.336
dqi_pc	0.108	-0.058	0.060	-0.010	-0.024	0.085	-0.066	0.563
staff_survey	-0.097	0.006	-0.269	0.127	0.087	-0.016	0.029	0.497
jundocpc	0.110	-0.175	0.037	-0.015	-0.043	0.033	-0.143	0.530
sick_rate	0.129	-0.082	0.030	0.117	0.018	-0.115	0.126	0.501
info_gov	0.067	-0.184	-0.098	-0.090	-0.051	-0.016	0.054	0.562

Results show a clear set of independent factors which correspond to the correlations found. These are once again:

- Inpatient satisfaction
- Waiting times
- Cancelled operations
- Readmission rates (for adults)
- Cancer waits
- Outpatient waits
- Financial balance

The fact that the CHI review emerges as a unique factor when it is included in the factor analysis, and the fact that it is highly correlated with the star ratings, suggests a very strong association between the CHI review and star rating and hence high implicit weighting in the algorithm used. This is because of the decision rules (or

Finsbury rules), for incorporating CHI's clinical governance review scores into star ratings (Commission for Health Improvement, 2003b).

The following table shows the very high correspondence between the star ratings and the CHI review, since the CHI review in essence drives the star rating (Jacobs and Smith, 2003). There is of course also likely to be some feedback effect from the previous year's ratings into the future CHI reviews which is likely to maintain this high association.

Table 27: Two-way table of relative frequency counts, percent and Pearson's chi-squared for CHI review against star rating, n=90

	chi_review				
pi_stars	1	2	3	4	Total
0	5	1	0	0	6
	100	1.72	0	0	6.67
	5.56	1.11	0	0	6.67
1	0	20	2	1	23
	0	34.48	18.18	6.25	25.56
	0	22.22	2.22	1.11	25.56
2	0	37	4	2	43
	0	63.79	36.36	12.5	47.78
	0	41.11	4.44	2.22	47.78
3	0	0	5	13	18
	0	0	45.45	81.25	20
	0	0	5.56	14.44	20
Total	5	58	11	16	90
	100	100	100	100	100
	5.56	64.44	12.22	17.78	100

This data analysis section has therefore explored the distributions of the variables included in the star ratings, their transformations and the new distributions produced, and finally the degree of collinearity between various sets of indicators.

6.2. Constructing a new composite indicator

The second part of the empirical analysis uses the above data to construct a new composite indicator in which the decisions made at each step can be explored and the impact they have on the robustness of the rankings of hospitals within the composite can be described.

This was achieved by performing Monte Carlo simulations and generating a simulated dataset. Monte Carlo simulation is a stochastic technique used to solve mathematical problems. The term 'stochastic' means that it uses random numbers and probability statistics to obtain an answer. Simulation means it is an analytical method meant to imitate a real-life system which is used when other analyses are too mathematically complex or too difficult to reproduce (Mooney, 1997).

Without the aid of simulation, the construction of the composite index will only reveal a single outcome. Monte Carlo simulation randomly generates values for uncertain variables over and over to simulate a model.

The term "Monte Carlo" comes from the name of the city in Monaco where the city's main attractions are casinos, which run games of chance such as roulette wheels, dice and slot machines; games which exploit the random behaviour of each game. The random behaviour in games of chance is similar to how Monte Carlo simulation selects variable values at random to simulate a model. For each uncertain variable (one that has a range of possible values), the possible values are defined with a probability distribution. The type of distribution selected is based on the conditions surrounding that variable. Distribution types include normal, log normal and so on. In this example, the variables all have a z score distribution ~N(0,1).

A simulation calculates multiple repetitions of a model by repeatedly sampling values from the probability distributions for the uncertain variables. Simulations can consist of as many repetitions as chosen. During a single repetition, a value is randomly drawn from the defined possibilities (the range and shape of the distribution) for each uncertain variable and selected to fit a probability distribution. This random draw process is then repeated many times. Each time a value is randomly drawn, it forms one possible solution (or in this case composite indicator). Together, these repetitions give a range of possible solutions, some of which are more probable and some less probable. Accuracy of this solution can be improved by increasing the number of repetitions.

This exercise uses 1000 repetitions and thus produces 1000 composite indices which produce a range of outcomes. These can then be used to produce 95 percent uncertainty intervals around the composite (truncating the data at the 2.5 and 97.5 percentiles). The interpretation of these intervals is that on 1000 repetitions, there is a 95 percent probability that the composite index will fall within the interval presented.

The analysis in this section first presents the construction of the composite without the simulations, explaining the choice of indicators, their transformations, their distributions and correlations. The report then proceeds to produce the simulated dataset (1000 composite indices) alongside the new composite and also shows the descriptive statistics and correlations for this simulated dataset. The simulated dataset is then used to produce uncertainty intervals around the new composite and the implications of this are explored. The empirical analysis then proceeds to use this composite (and the uncertainty intervals) to examine a number of important issues, namely:

1. Decomposing the variation on the performance indicators
2. Introducing different weighting systems
3. Transforming the indicators to a categorical scale
4. Introducing decision rules to construct the composite

These issues are essentially some of the key features of the star rating system, although the issues are generic to many other composites. The analysis shows the conditions under which these issues are likely to impact on the robustness of the rankings of units within a composite.

6.2.1. The chosen indicators

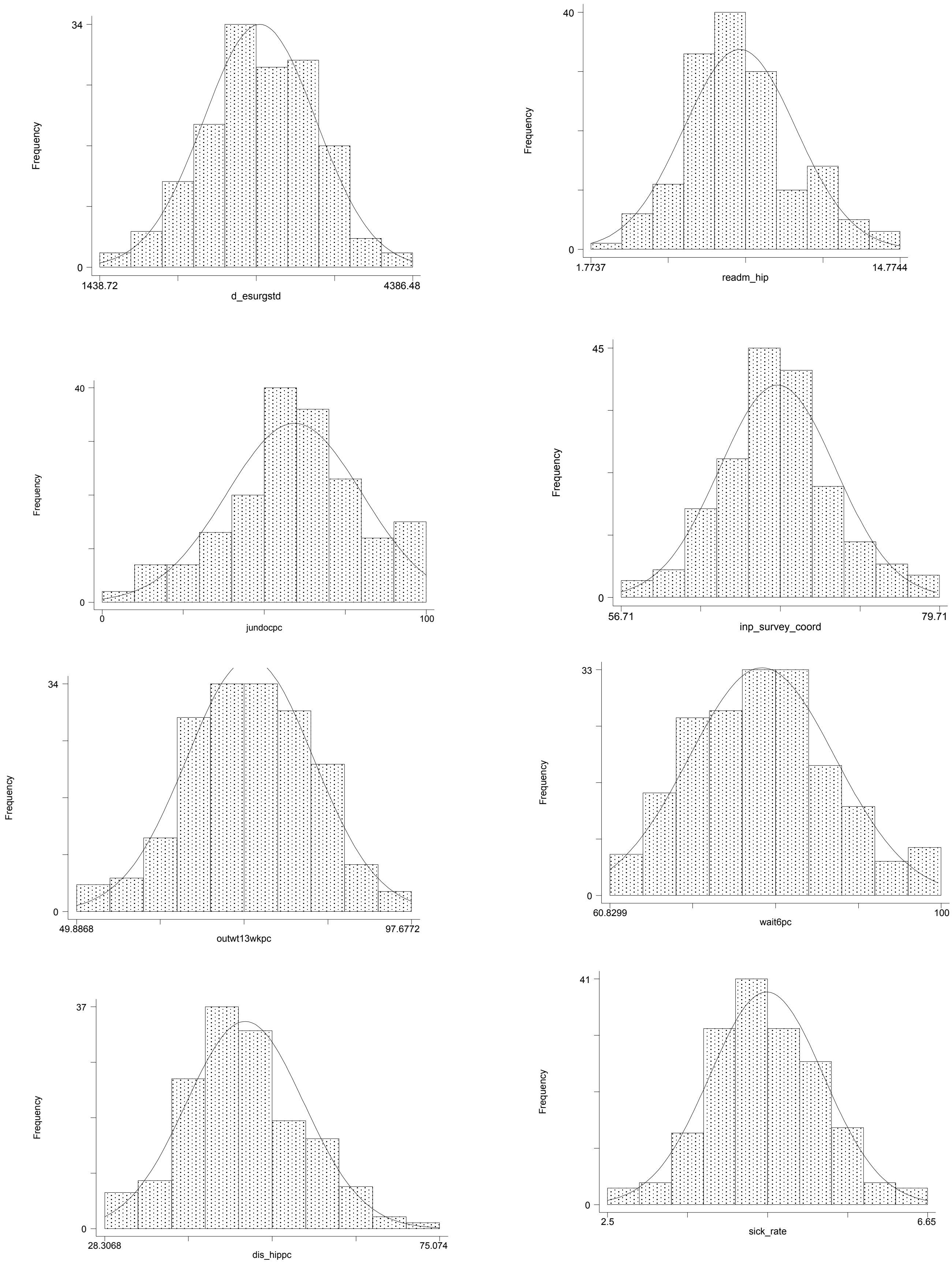
Of the 38 indicators in the Star ratings, 10 variables were selected for the example. These were chosen on the basis of the data analysis in the first part of the empirical analysis. Examining the correlations and the factors produced in the factor analysis, indicators were chosen that seemed to represent independent and important performance dimensions. Thus variables were chosen that did not have very high collinearity, whilst trying to include most of the independent dimensions emerging from the factor analysis. Variables were also only included if they exhibited approximately continuous normal distributions. Thus threshold type variables reflecting service standards were excluded in preference of balanced scorecard type variables which pick up differential performance.

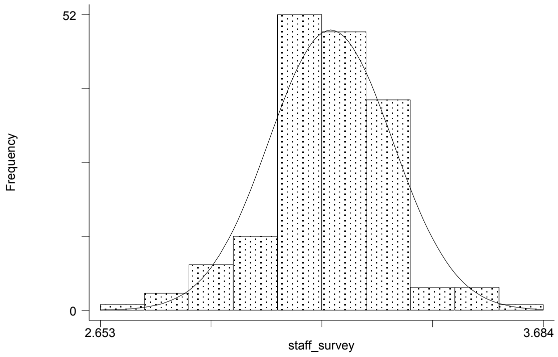
The following figures show the distributions for the ten variables chosen. They include:

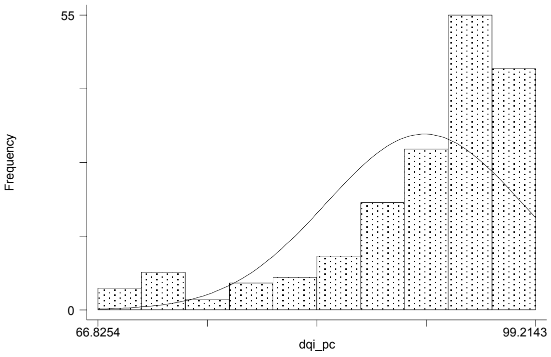
1. death rates following emergency surgery
2. readmission rates for hip fracture
3. junior doctors working hours
4. inpatient satisfaction with coordination of care
5. thirteen week outpatient waits
6. six month inpatient waits
7. discharge home following hip fracture
8. sickness absence rate for staff
9. staff satisfaction
10. data quality

The figures highlight the approximately normal distributions of all the variables.

Figure 12: Histograms and normal density distributions for 10 variables chosen for new composite







These ten variables were then transformed (standardised) to a z score to each have a distribution of mean zero, unit variance (0,1). Since these variables were all approximately normal and there were no outliers, this standardisation is a reasonable choice of transformation. Since the Monte Carlo simulations can be problematic in the face of missing data, all units with missing data were excluded from the analysis, reducing the sample to 117 hospital trusts.

The following table gives the descriptive statistics for the transformed variables. As mentioned, since the raw data was not first re-scaled, the indicators will not have exactly the same range (max – min). The transformed variables also mostly appear to be approximately normal and symmetrical.

In order to construct the new composite index, three of the variables needed to be re-interpreted as 'more is better' type variables and hence were simply converted into negative variables. These were death rates following emergency surgery, readmission rates for hip fracture and sickness absence rates for staff. Following the standardisation described above, all ten variables could then simply be summed together in a linear fashion to form the new composite.

The descriptive statistics for the new composite indicator are also shown in the table. As expected, the standard deviation for the composite will also be much larger than for each of the underlying indicators. The hospitals could then also be ranked on the new composite from 1-117.

Table 28: Descriptive statistics of standardised variables and new composite

variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
mind_esurgstdst	117	0	-0.021	1	-2.669	2.068	-0.009	2.520	0.516
dis_hippcst	117	0	-0.053	1	-2.284	3.200	0.322	3.180	0.258
wait6pcst	117	0	-0.012	1	-2.257	2.637	0.231	3.011	0.530
outwt13wkpcst	117	0	-0.009	1	-2.414	2.739	-0.011	2.888	0.998
inp_survey_coordst	117	0	0.043	1	-3.009	2.494	-0.162	2.892	0.751
jundocpcst	117	0	0.073	1	-2.398	2.157	-0.287	2.650	0.326
minsick_ratest	117	0	0.089	1	-2.739	2.744	-0.306	3.144	0.301
minreadm_hipst	117	0	0.143	1	-2.697	2.551	-0.422	3.104	0.135
staff_surveyst	117	0	-0.089	1	-3.878	2.640	-0.669	4.576	0.002
dqi_pcst	117	0	0.298	1	-3.578	1.205	-1.505	5.099	0.000
composite	117	0	0.634	3.361	-9.679	7.503	-0.654	3.269	0.019

The following table shows the correlations between the 10 standardised variables to form the new composite as well as with the new composite. Generally none of the indicators have correlations greater than ±0.3.

Table 29: Correlations between 10 standardised variables to form new composite, n=117

	min	dis_hippc	wait6pc	outwt13w	inp_survey	jundocpc	min	min	staff_survey	dqi_pc	composite
	d_esurgstd	st	st	kpc	_coord	st	sick_rate	readm_hip	st	st	
	st			st	st		st	st			
mind_esurgstdst	1.0000										
dis_hippcst	-0.0063	1.0000									
wait6pcst	-0.0174	-0.0982	1.0000								
outwt13wkpcst	0.0399	-0.2042	0.2527	1.0000							
inp_survey_coordst	0.2424	0.0339	0.0936	0.2189	1.0000						
jundocpcst	-0.0367	-0.1242	0.0918	-0.0827	0.1045	1.0000					
minsick_ratest	0.0680	0.1159	-0.1765	0.0780	-0.0376	-0.3095	1.0000				
minreadm_hipst	0.1142	-0.1129	-0.1611	-0.0813	0.1110	-0.0264	0.2264	1.0000			
staff_surveyst	0.0716	0.0481	0.1195	0.0936	-0.0497	-0.1582	0.1154	-0.1045	1.0000		
dqi_pcst	-0.1686	-0.1951	0.3087	0.1570	0.1237	0.0321	-0.1141	0.0072	0.0460	1.0000	
composite	0.3889	0.1360	0.4204	0.4379	0.5476	0.1460	0.2874	0.2894	0.3516	0.3561	1.000

6.2.2. Monte Carlo simulations

Simulations were then performed by drawing samples from a multivariate normal distribution to emulate the above 10 indicators and their correlations with one another. Thus by sampling values from the probability distributions for the ten variables with the same zero means and the same covariance matrix, 1000 random versions of the above dataset could be reproduced. The dataset was generated for data with mean zero and variance one ~ N(0,1) for each sample (the same as the original 10 indicators). Each sample was drawn with the above correlation structure (of the underlying standardised variables), for 1000 replications. For each sample a new composite was constructed and a ranking based on the new composite.

The following table shows the descriptive statistics for the simulation sample. For *n* now equal 117,000, the indicators have approximately mean zero unit variance. However the range is a bit wider than for the original composite since there is greater uncertainty around the indicators and across the 1000 draws. Once again the indicators are on average approximately normal.

Table 30: Descriptive statistics of standardised variables from simulations (1000 replications)

Variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
mind_esurgstdst	117000	-0.0012	-0.0026	1.0011	-4.6569	4.7106	0.619	0.554	0.7416
dis_hippcst	117000	0.0026	0.0042	1.0015	-4.3791	4.2569	0.360	0.629	0.5854
wait6pcst	117000	-0.0010	0.0000	0.9984	-4.1414	4.3906	0.812	0.864	0.9581
outwt13wkpcst	117000	-0.0035	-0.0008	1.0021	-3.9339	4.4132	0.179	0.576	0.3475
inp_survey_coordst	117000	-0.0032	-0.0021	0.9987	-4.5566	4.1575	0.353	0.404	0.4589
Jundocpcst	117000	-0.0029	-0.0033	0.9967	-4.5327	4.4441	0.444	0.037	0.0856
minsick_ratest	117000	0.0027	0.0035	0.9999	-4.3413	4.2740	0.448	0.880	0.7408
minreadm_hipst	117000	0.0050	0.0041	1.0007	-5.5077	4.8468	0.074	0.038	0.0235
staff_surveyst	117000	0.0026	0.0033	0.9991	-4.5720	4.2472	0.972	0.452	0.7524
dqi_pcst	117000	-0.0023	-0.0015	0.9989	-4.4641	4.4311	0.498	0.328	0.4926

The following table acts as a test essentially to ensure that the covariance matrix structure and the correlations between the 10 standardised indicators across the 1000 replications approximately emulates the original covariance matrix structure. The correlations, compared to Table 29, are indeed very similar.

Table 31: Correlations between 10 standardised variables from simulations (1000 replications), n=117000

	min			outwt13w	inp_survey		min	min		
	d_esurgstd	dis_hippc	wait6pc	kpc	_coord	jundocpc	sick_rate	readm_hip	staff_survey	dqi_pc
	st	st	st	st	st	st	st	st	st	st
mind_esurgstdst	1.0000									
dis_hippcst	-0.0108	1.0000								
wait6pcst	-0.0164	-0.0980	1.0000							
outwt13wkpcst	0.0390	-0.2067	0.2499	1.0000						
inp_survey_coordst	0.2464	0.0354	0.0900	0.2179	1.0000					
jundocpcst	-0.0363	-0.1170	0.0899	-0.0855	0.1079	1.0000				
minsick_ratest	0.0676	0.1149	-0.1790	0.0765	-0.0386	-0.3088	1.0000			
minreadm_hipst	0.1174	-0.1167	-0.1594	-0.0794	0.1112	-0.0251	0.2239	1.0000		
staff_surveyst	0.0718	0.0470	0.1204	0.0957	-0.0501	-0.1595	0.1200	-0.1076	1.0000	
dqi_pcst	-0.1686	-0.1903	0.3066	0.1515	0.1231	0.0296	-0.1147	0.0078	0.0448	1.0000

In order to construct the new composite, the original scores from each performance indicator for each hospital are then added to the simulated dataset to obtain the following descriptive statistics. As can be seen the standard deviation for each indicator increases to approximately 1.4 and the range increases (for the same reasons) from ±4.5 in Table 30 to ±6 in Table 32. The standard deviation on the composite is also commensurately larger at 4.8.

Table 32: Descriptive statistics of standardised variables and new composite from simulations (1000 replications)

variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
mind_esurgstdst	117000	-0.0012	-0.0080	1.4138	-5.9369	5.6063	0.904	0.000	0.0000
dis_hippcst	117000	0.0026	-0.0318	1.4157	-5.6570	7.0279	0.000	0.000	0.0000
wait6pcst	117000	-0.0010	-0.0134	1.4136	-5.7476	5.8587	0.000	0.351	0.0000
outwt13wkpcst	117000	-0.0035	0.0043	1.4125	-5.6276	6.0684	0.506	0.047	0.1122
inp_survey_coordst	117000	-0.0032	0.0149	1.4068	-6.4436	5.4590	0.000	0.000	0.0000
jundocpcst	117000	-0.0029	0.0327	1.4072	-5.6347	6.3681	0.000	0.000	0.0000
minsick_ratest	117000	0.0027	0.0346	1.4105	-5.8227	5.7476	0.000	0.279	0.0000
minreadm_hipst	117000	0.0050	0.0499	1.4143	-6.2602	5.6764	0.000	0.009	0.0000
staff_surveyst	117000	0.0026	0.0409	1.4130	-7.0770	5.9946	0.000	0.000	
dqi_pcst	117000	-0.0023	0.1227	1.4118	-6.7770	5.4617	0.000	0.000	
composite	117000	-0.0013	0.1971	4.7504	-21.5216	17.6654	0.000	0.000	0.0000

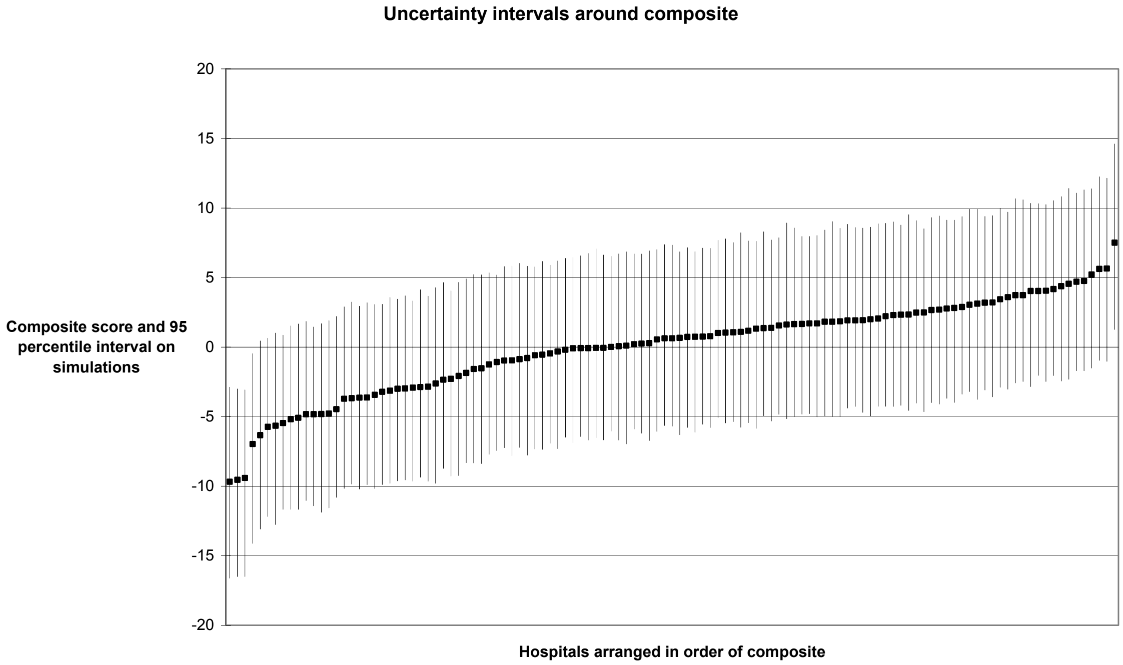
Once again the correlations for each of the 1000 samples approximately match the correlation structure of the underlying 10 standardised variables from the original composite, as does the correlation matrix for the pooled dataset for all 1000 samples given in the table below.

Table 33: Correlations between 10 standardised variables from simulations with composite (1000 replications), n=117000

	min d_esurgstd st	dis_hippc st	wait6pc st	outwt13w kpc st	inp_survey _coord st	jundocpc st	min sick_rate st	min readm_hip st	staff_survey st	dqi_pc st	composite
mind_esurgstdst	1.0000										
dis_hippcst	-0.0071	1.0000									
wait6pcst	-0.0160	-0.0985	1.0000								
outwt13wkpcst	0.0390	-0.2052	0.2517	1.0000							
inp_survey_coordst	0.2452	0.0343	0.0934	0.2189	1.0000						
jundocpcst	-0.0346	-0.1221	0.0902	-0.0826	0.1064	1.0000					
minsick_ratest	0.0686	0.1169	-0.1773	0.0769	-0.0361	-0.3075	1.0000				
minreadm_hipst	0.1163	-0.1136	-0.1615	-0.0813	0.1137	-0.0252	0.2242	1.0000			
staff_surveyst	0.0713	0.0490	0.1221	0.0948	-0.0463	-0.1604	0.1182	-0.1063	1.0000		
dqi_pcst	-0.1709	-0.1940	0.3077	0.1552	0.1240	0.0293	-0.1143	0.0073	0.0481	1.0000	
composite	0.3901	0.1374	0.4197	0.4360	0.5499	0.1456	0.2884	0.2895	0.3542	0.3542	1.0000

The following figure shows the ranking of the 117 hospitals on the new composite shown in order from the worst to the best performer (the dark dots). If the simulated data were not produced one might conclude that the performance of the best hospital on the composite appears to be significantly better than the performance of the worst hospital (comparing –9.7 and 7.5 from Table 28). However, when the 95 percent uncertainty intervals are produced around the composite the conclusions appear quite different. Around each of the dark dots (each hospital unit) a line shows the range (max – min) in which the composite for this hospital could potentially fall 95 percent of the time. Thus drawing from a multivariate normal distribution where the indicators have the same distributions ~ N(0,1) and the same correlations exist between them, a composite could potentially be constructed which falls within this range for each hospital. Since these intervals essentially overlap over almost the entire range of the composite, one cannot be certain that (say) the hospital ranked 10 is necessarily performing better on the composite constructed for (say) the hospital ranked 50, using these 10 indicators.

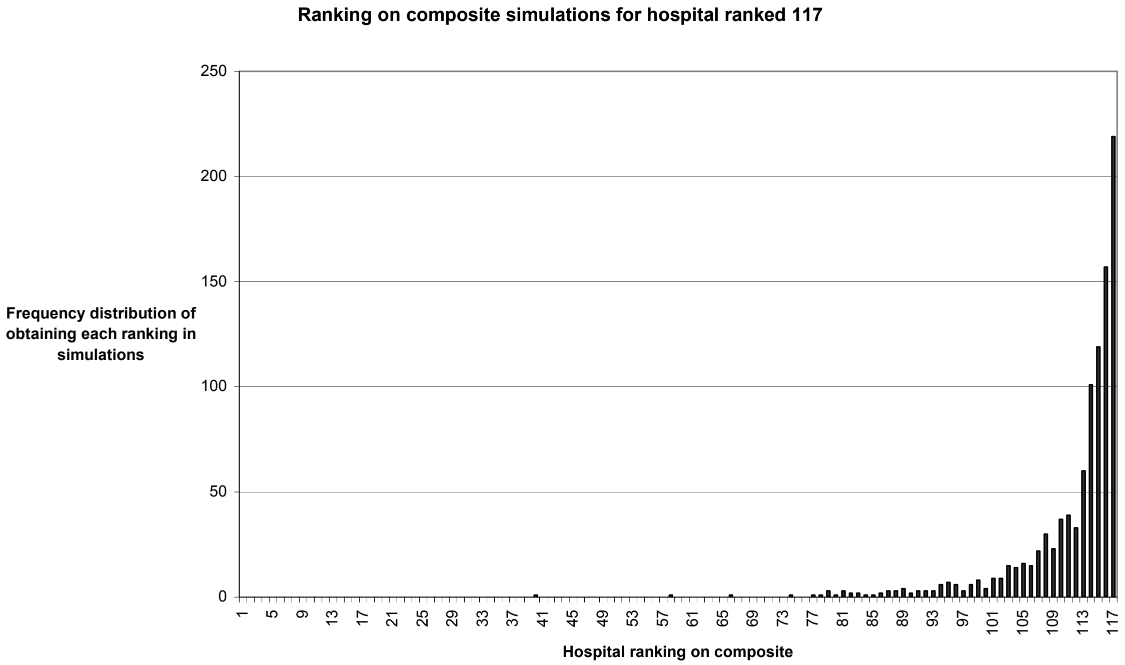
Figure 13: Uncertainty intervals around composite score using simulations



It is also of course the case that some of the variation that exists around the composite (and the underlying 10 indicators used to construct the composite) will be random variation, the result of measurement error or sampling error or simply natural variation that exists within any distribution. If it were possible to decompose the variation into this element, as opposed to the true performance variation, one may be able to say with greater certainty that the hospital ranked 10 is performing better on the composite than the hospital ranked 100 (assuming that the uncertainty intervals of these hospitals no longer overlap), since the variation around the composite could be reduced. This issue is explored in the following section.

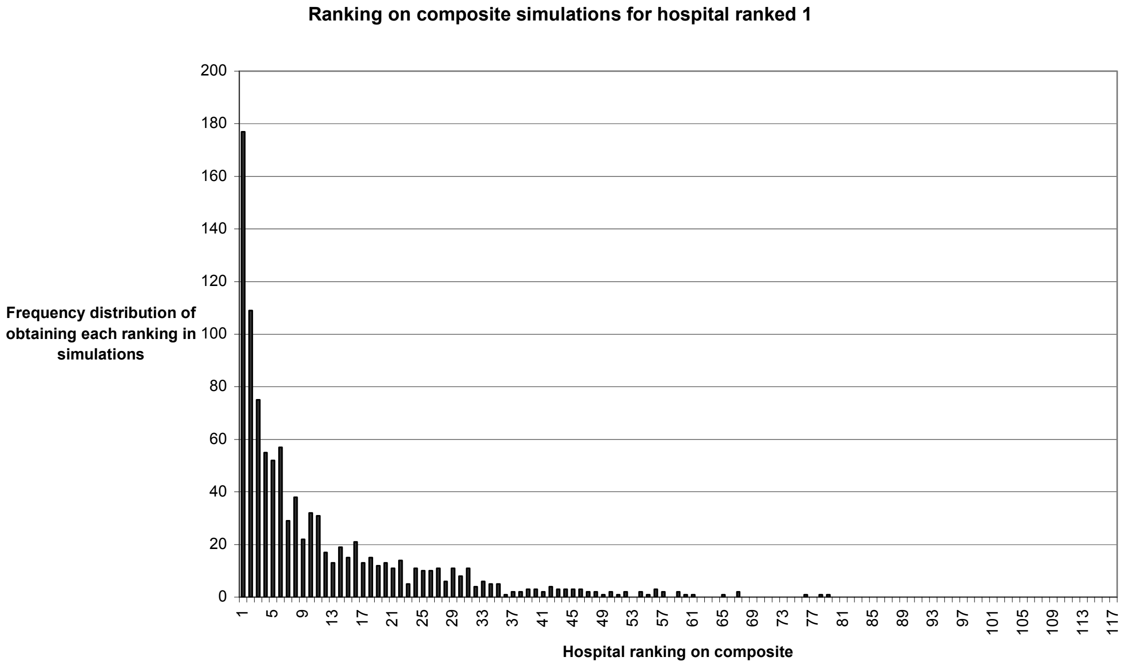
Since the concern is with the robustness of the rankings of hospitals on the composite, the following two figures show the frequency distribution for the hospital ranked 117 (worst) and 1 (best) on the original composite respectively, across the 1000 simulations. These show that whilst the frequency with which these hospitals get ranked at the bottom and top of the distributions respectively are certainly higher, there is a probability that they could be ranked right down to the middle or other ends of the distribution, almost jumping two-thirds of the league table in either direction.

Figure 14: Frequency distribution of ranking on composite score using simulations for worst ranked hospital



There seems from these figures to be somewhat greater stability in the ranking of the worst hospital than there is in the best ranked hospital.

Figure 15: Frequency distribution of ranking on composite score using simulations for best ranked hospital



6.2.3. Random variation on performance indicators

As mentioned before, while performance on the underlying indicators may be identifiable to some extent, much of the variation in performance across the 117 hospitals on any particular indicator will be due to a number of indecipherable and unpredictable events such as measurement error or sampling error or simply normal random variation. When comparisons are therefore made between hospitals and over time, account must be taken of the variability in performance due to such random events. In practice this means one must know something about how a purely random process generates a distribution of events. As mentioned in the discussion about the types of indicator variables, many of these counts of events (such as deaths or readmissions) are assumed to emanate from a Poisson distribution which assumes that the events are only subject to random variation. However, it is highly unlikely that any of these indicators will have random variation as the only source of variation. Being able to disentangle these different sources of variation in the clinical indicators (which will translate through into the construction of the composite) will make the use of the uncertainty intervals much more useful since much tighter intervals can be obtained if some of the variation can be taken out. The objective is therefore to obtain an estimate of the variation on each indicator which is due to random variation (measurement error and natural variation).

There is no agreed methodology to do this and as such the best method was deemed to be exploiting the *within* and *between* variation that exists when longitudinal data is available. Although the star ratings having been running for three years, the changes in the inclusion and exclusion of certain variables means that for some indicators longitudinal data was not available. However, for other indicators, more than three years of data were available. The longer the time series available, the better estimate can be made of the degree of random variation *within* hospitals over time and *between* different hospitals.

In order to get an estimate of the random variation on each performance indicator, a fixed effects panel data regression was run on each of the variables which have data for more than 1 year (7 of the 10 variables). Fixed effects regressions are essentially the same as ordinary least squares regressions when indicator (dummy) variables are included for each of the hospitals. The fixed-effects model is:

$$y_{it} = a + x_{it}\, b + v_i + e_{it} \tag{18}$$

where v_i are the hospital-specific fixed effects and year (time) dummies are included on the right-hand side.

The results for each of these regressions produce an estimate of the within variation explained by the hospital specific effects and the year dummy variables. This R-squared within variation on each of the indicators is then used as an estimate to explain the proportion of random variation which exists on each of the indicators. The within variations for each of the indicators were as follows:

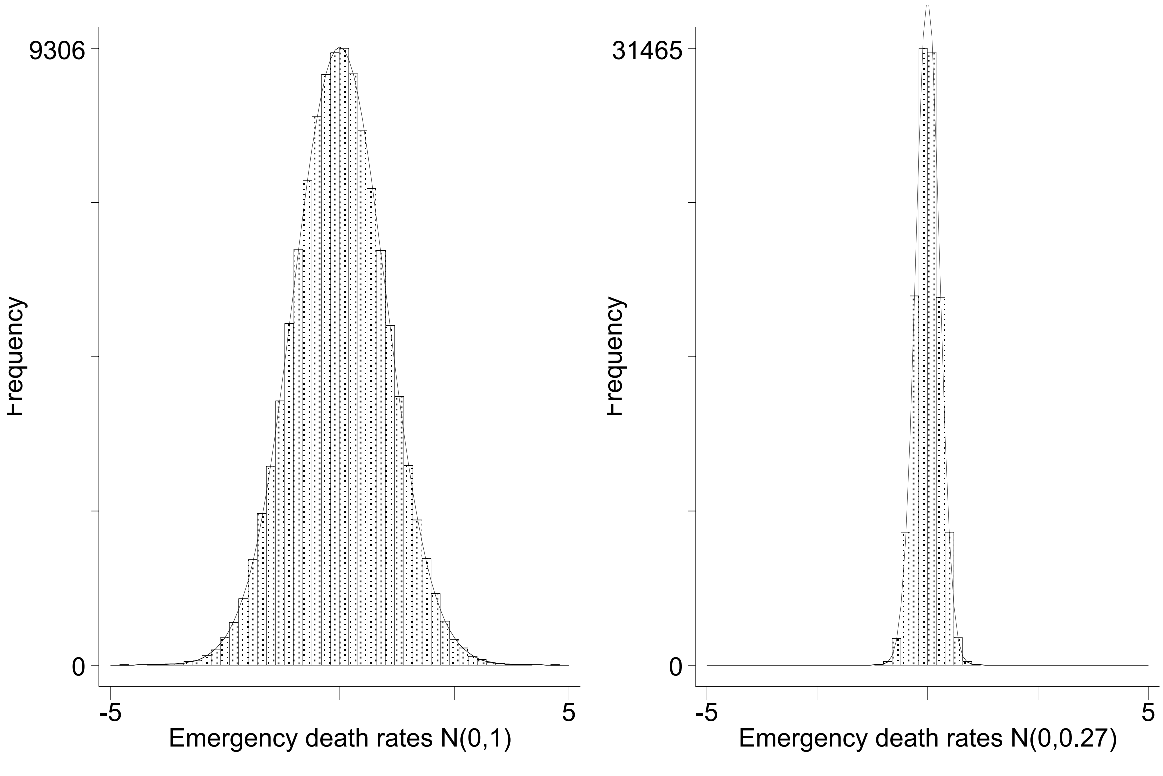
- 27 percent for emergency death rates
- 25 percent for 6 month inpatient waits
- 20 percent for readmissions following hip fractures
- 14 percent for discharges following hip fractures

- 13 percent for 13 week outpatient waits
- 4 percent for data quality
- 2 percent for sickness absence rates
- 15 percent for patient satisfaction with coordination (assumed)
- 15 percent for junior doctors hours (assumed)
- 15 percent for staff satisfaction (assumed)

For the last three variables for which there was no panel data available, an average figure of 15 percent was assumed. There is therefore most variation around the death rates and waiting times indicators and least variation around indicators on data quality and sickness absence which is to be expected since greater natural variation is likely to exist around clinical indicators compared to measures such as data quality which are potentially subject to greater managerial control and mediation and less subject to random events. These results accord with previous research findings (Hauck *et al*, 2003).

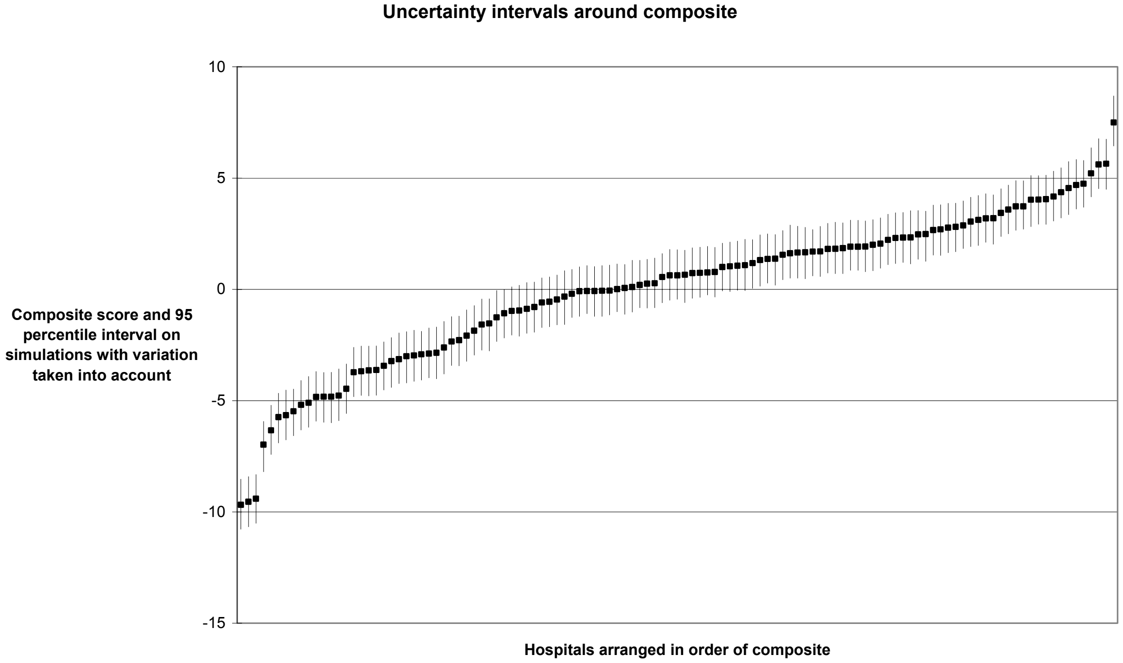
These estimates on the proportion of random variation around the individual performance indicators were then used to adjust the variation around each of the indicators to the estimate obtained. Thus the standard deviation on each of the indicators now essentially reduces to the 'random' within variation which remains on each of the indicators as listed above. Thus in the following figure the variation around the distribution for death rates reduces from ~ N(0,1) to ~ N(0,0.27). This exercise was carried out for each of the ten variables.

Figure 16: Example of variance around emergency death rates with 1000 observations with distribution N(0,1) and N(0,0.27)



The variables with the reduced variation were now used to re-construct the uncertainty intervals around the composite indicator. While the black dots of the original composite remain unchanged in the following figure, the variation around the composite has now shrunk dramatically as a result of the reduction in variation around the underlying indicators to the proportion of variation due to random events.

Figure 17: Uncertainty intervals around composite score using simulations with random variation taken into account



Thus the ability to decompose the variation around performance indicators is a crucially important exercise. The interpretation of the results of the composite now change dramatically and one can now say with greater certainty that hospital 10 is performing much better than say hospital 100. The importance of producing uncertainty intervals as a means of communicating the range of estimates that can be obtained for the composite index, is highlighted as a worthwhile analytical practice since it can dramatically change conclusions about performance.

6.2.4. Introducing changes in weights

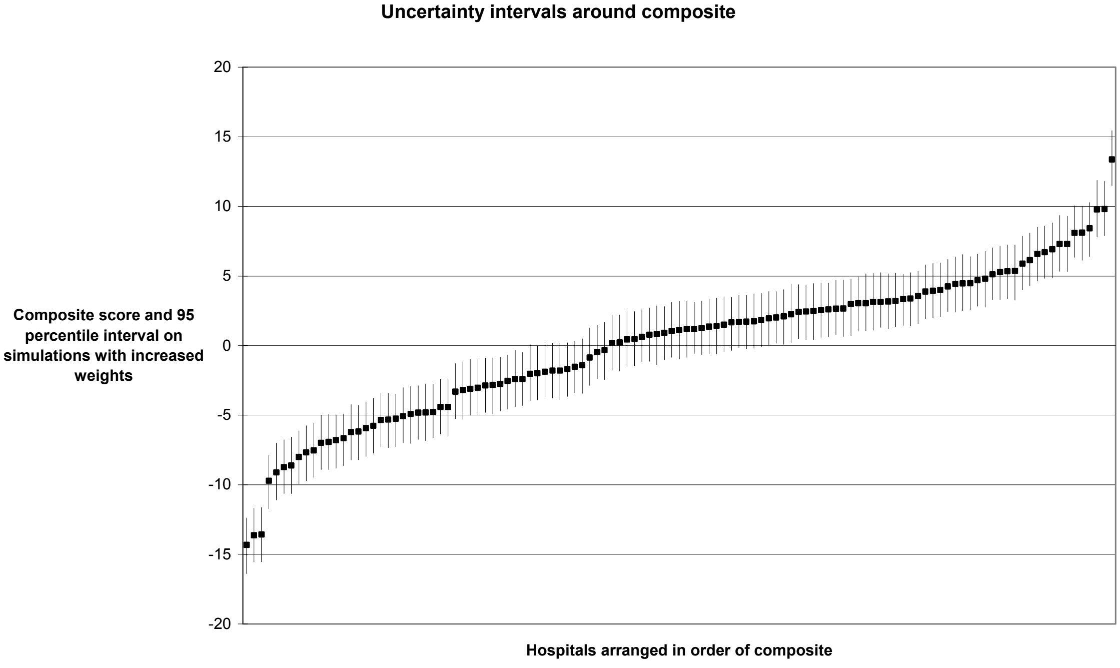
This section explores the sensitivity of the rankings on the composite index to changes in weighting systems. The composite index with the narrower uncertainty intervals given above, is used to explore changes in the weights applied to the underlying 10 indicators (or sub-sets). When the weight of a particular variable is changed, the standard deviation of that particular variable changes according to the new weight, thus if the weight is (say) doubled (×2), the standard deviation also doubles (1×2), assuming a z score (0,1) distribution. The composite still has zero mean but the standard deviation (and variance) also increases, although it has a larger standard deviation than the underlying indicator. Of course the opposite happens if the weight of a particular variable is reduced, its standard deviation reduces by the chosen amount, and the variance on the composite also decreases.

A large number of scenarios were tested for changes in weighting structures, in particular exploring the relationship between the correlation of variables and changes in the weights applied to them. Two potential policy relevant examples of hypothetical scenarios are shown below.

In the following example the weights on the outcome and waiting times variables (5 of the 10 variables) are doubled. These are death rates following emergency surgery, readmission rates for hip fracture, discharge home following hip fracture, thirteen week outpatient waits, and six month inpatient waits. Thus setting aside for now the difficulties discussed in eliciting preferences and the important considerations around whose preferences are in fact elicited, a possible scenario is that greater emphasis is placed on these outcome and waiting times objectives by (say) policy-makers.

The following figure shows the uncertainty intervals around the new composite following the above increase in weights.

Figure 18: Uncertainty intervals around composite score using simulations with an increase in weights on five variables

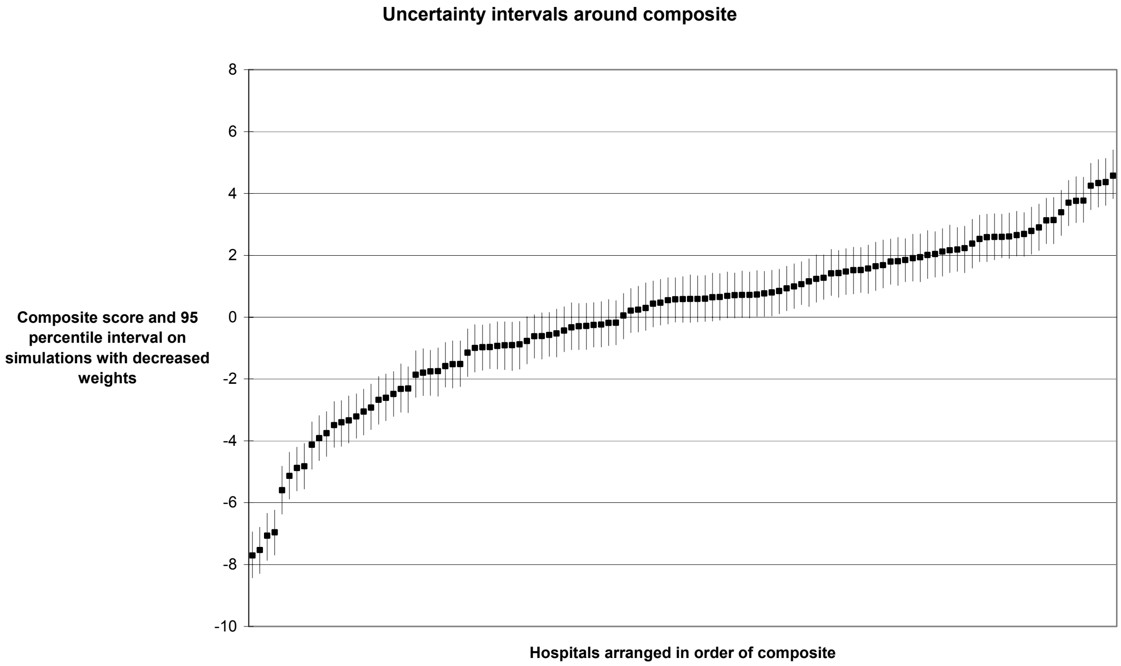


As can be seen the composite now stretches over a wider range (-14.3 to 13.4) from (-9.7 to 7.5). The correlation with the original composite index is 0.97 and between the two sets of hospital rankings is 0.96.

The best ranked hospital from the original composite index now receives a score of 13.4 and is ranked 1 whereas the worst ranked hospital receives a score of –13.6 and is ranked 116 (up from 117). The largest change in rankings for an individual unit is 34 places while the average change in ranking from the original composite is 7 places.

In the following example the same outcome and waiting times variables (the same 5 of the 10 variables), have their weights halved instead of doubled. The results are shown in the following figure.

Figure 19: Uncertainty intervals around composite score using simulations with a decrease in weights on five variables

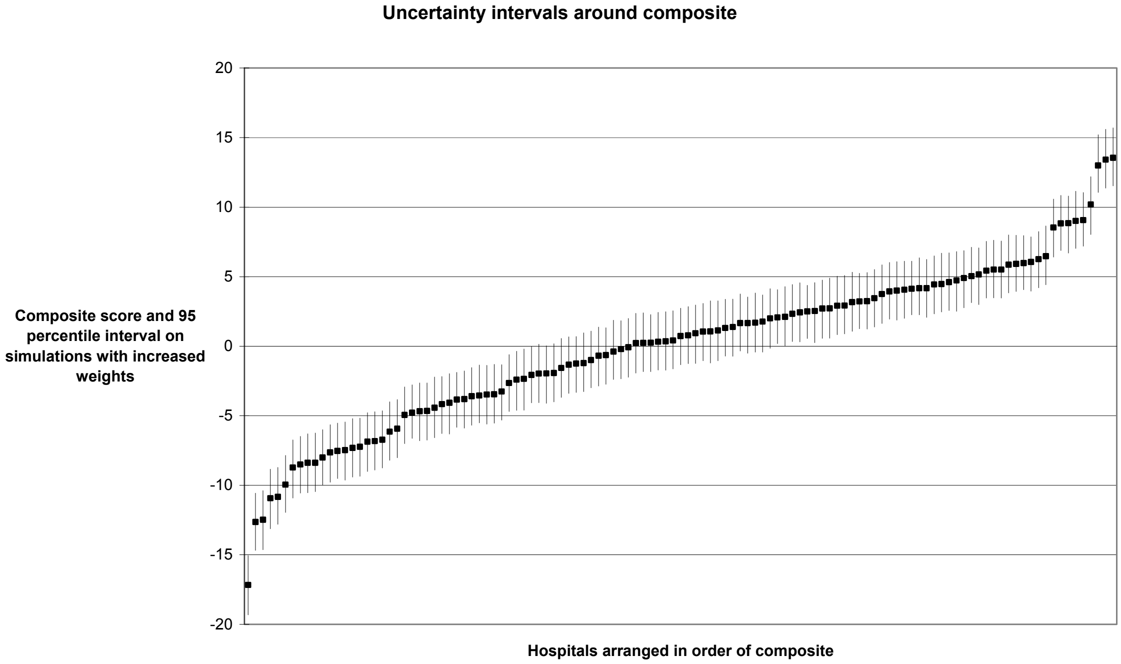


The range predictably shrinks to (-7.7 to 4.6). The correlation with the original composite index is 0.97 and between the two sets of hospital rankings is 0.96.

The best ranked hospital from the original composite index now receives a score of 4.6 and is again ranked 1 whereas the worst ranked hospital receives a score of –7.7 and is again ranked 117. The largest change in rankings for an individual unit is 39 places, a third of the league table, while the average change in ranking from the original composite is again 7 places.

In the following alternative two scenarios, the weights on the two waiting times variables (2 of the 10 variables) are increased threefold (six month inpatient waits and thirteen week outpatient waits). Since waiting times are such a policy priority, this would seem a plausible hypothetical scenario. Once again the dataset with the reduced variation is used to test the change in weights. The following figure shows the new composite index.

Figure 20: Uncertainty intervals around composite score using simulations with an increase in weights on two variables

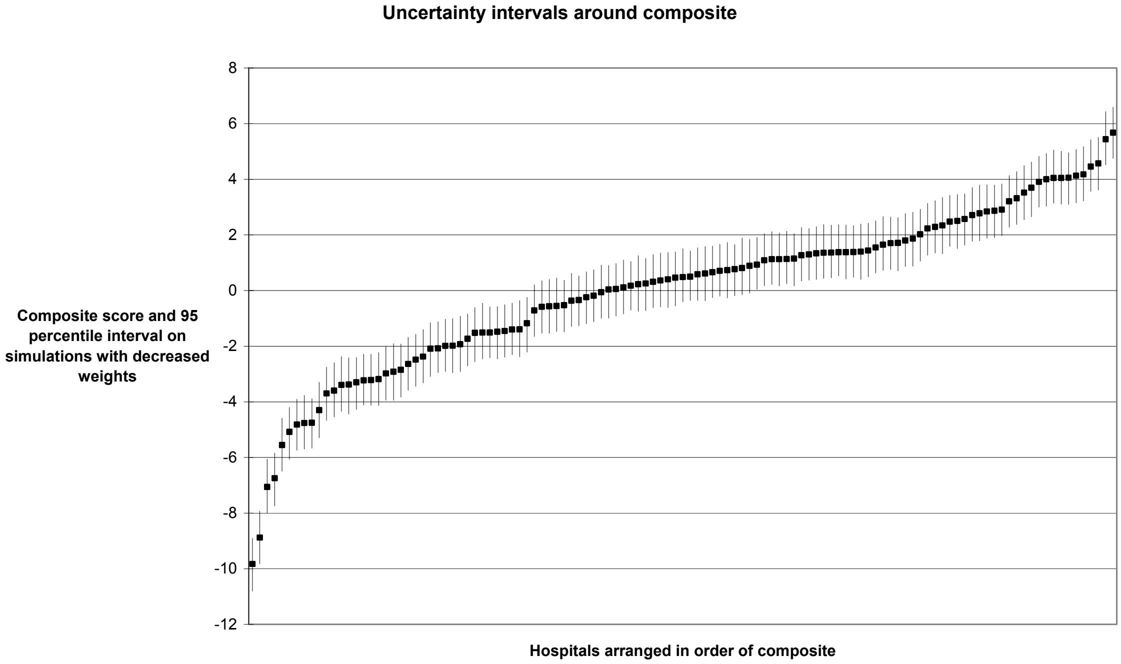


Once again, the increase in weights on individual indicators will increase the standard deviation and variance on each indicator and the variance on the resultant composite. The slightly wider uncertainty intervals can be seen in the above figure. As a result, once again the range for the new composite is wider, since there is more variation in the new composite (from –17.2 to 13.5). The correlation with the original composite index is 0.89 and between the two sets of hospital rankings is 0.88. There is therefore a bigger change in the correspondence with the original set of rankings compared to the first example shown.

The best ranked hospital from the original composite index now receives a score of 13.4 and is ranked 2 whereas the worst ranked hospital receives a score of –17.2 and is again ranked 117. The largest change in rankings for an individual unit is 54 places, nearly half the league table, while the average change in ranking from the original composite is 13 places or a decile, a pretty significant movement.

In the following example the weights on the same waiting times variables (2 of the 10 variables) are instead reduced to a third, as opposed to being increased threefold. The results are shown in the following figure.

Figure 21: Uncertainty intervals around composite score using simulations with a decrease in weights on two variables



The range of scores reduces to (-9.8 to 5.7). The correlation with the original composite index is 0.95 and between the two sets of hospital rankings is 0.92.

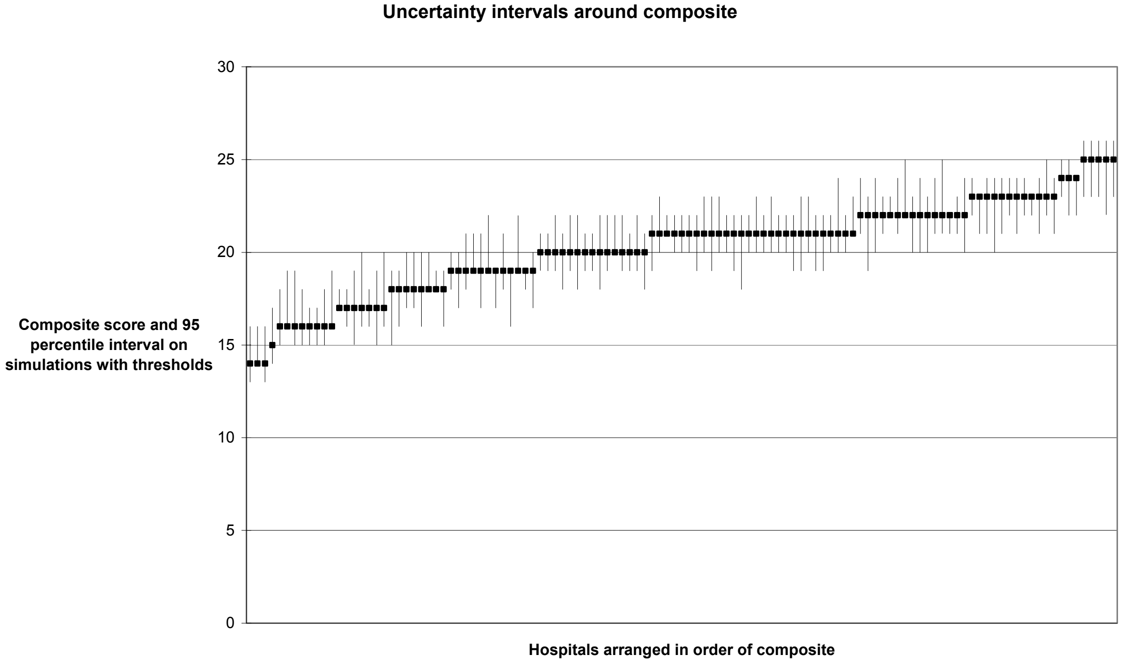
The best ranked hospital from the original composite index now receives a score of 5.4 and is ranked 2 whereas the worst ranked hospital receives a score of –7.1 and is ranked 115. The largest change in rankings for an individual unit is 45 places while the average change in ranking from the original composite is 10 places, again suggesting potentially large changes for individual units.

As suggested by these examples, the weighting structure can indeed materially affect the rankings of hospitals on the composite. Predictably, the ranks change the most for units when weights for indicators on which they perform exceptionally (well or poorly) are increased or decreased. Changing the weight structure and the impact this ultimately has on rankings, is related to the degree of correlation between the underlying indicators. This relates to the formal relationship given in equations (12) and (13) which show that the change in weights is intimately linked to the covariance between indicators.

6.2.5. Transforming into categorical variables

This section explores the impact of the transformation of the indicators into categorical variables, on the rankings of the hospitals. As mentioned the choice of thresholds used to transform the variables are often quite subjective and arbitrary and often vary across the different indicators. In this example, the same thresholds are applied to all 10 indicators. Each of the 10 standardised variables are transformed into a categorical variable representing thresholds of performance. There are 3 categories, those with a score above 0.5 achieve a 3, those with a score below –0.5 achieve a 1 and the rest in between score a 2. This partitions the indicators into approximately equal portions on the variables. The composite was constructed in the same way as before, as a simple linear summation of the underlying 10 indicators, now in categorical form. The following figure shows the new composite and the uncertainty intervals surrounding it.

Figure 22: Uncertainty intervals around composite score using simulations with thresholds



The range of the new composite (14 to 25) is essentially quite narrow since only three categories were chosen.

The correlation with the original composite index is 0.91 and between the two sets of hospital rankings is 0.91 (though the composite now takes a different form and many of the rankings are also shared since there are fewer permutations with categorical data).

The best ranked hospital from the original composite index now receives a score of 25 and is ranked 1 (several ranks are shared however) whereas the worst ranked hospital receives a score of 13 and is ranked 115 (also shared). The largest change in rankings for an individual unit is 43 places while the average change in ranking from the original composite is 13 places, around a decile of the distribution.

Visually it is clear that the nature of the uncertainty interval looks quite different under categorical variables to the previous composite. For many of the hospitals the composite score (dark dot) is at one end of the range of possible composite scores (the vertical line), rather than being somewhere in the middle of the distribution. This is because each of the distributions for each hospital are now categorical in nature. The impact on rankings produced by the changes in the thresholds is therefore understandably potentially large.

6.2.6. Introducing decision rules

In this final section, the decision rules that are often employed in the construction of composites were simulated. The purpose of this was to try to emulate the algorithm process used to construct the Star ratings, though not the actual algorithm itself, simply the process involved. Thus a new composite index was constructed, like the star ratings, with four categories from zero to three based on the above categorical variable transformations. The rules were chosen based on the fact that the variables were transformed to categorical variables on a scale of 1 to 3. A number of different combinations and permutations of decision rules were tried. The example below shows one hypothetical (though plausible) set of rules:

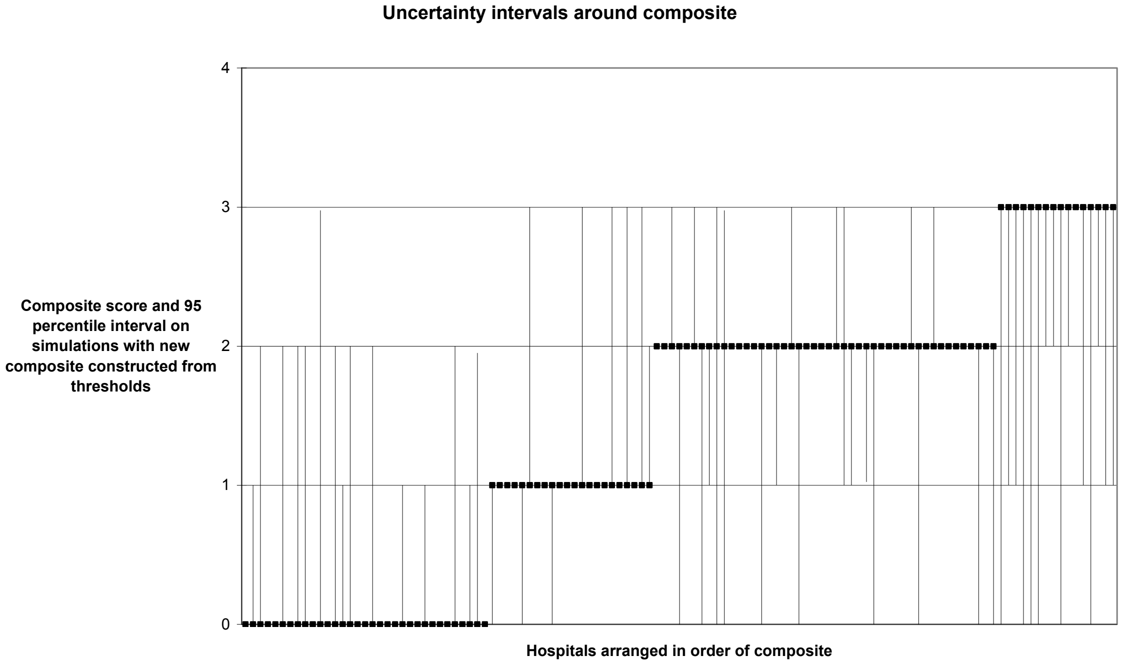
1. Hospitals receive a three star if they achieve either a two or a three on all three of the outcome variables: readmission rates for hip fracture, death rates following emergency surgery, and discharge home following hip fracture; and on the two waiting times variables: thirteen week outpatient waits, and six month inpatient waits.
2. Hospitals receive a zero star if they achieve a one or two on all of the three variables: inpatient satisfaction, staff satisfaction, and junior doctors working hours.
3. The rest of the hospitals receive a one star.
4. Those hospitals which receive a two star are distinguished from those receiving a one star if they also achieve a three score on any of the two variables: data quality and sickness absence rate.

Applying the following rules of the algorithm in the order above effectively places a set of implicit weights on the variables which are used to dictate the thresholds for best and worst performance. Thus the five variables in rule 1 will implicitly receive a higher weighting, which will therefore impact on the rankings of hospitals.

In the different scenarios that were tested, it is clear that subtle and highly subjective changes to the decision rules can dramatically impact on how many hospitals end up in each category. These decision rules were chosen so as to try to ensure an approximately equal number of hospitals in each group. However, the analyst could easily change the number of hospitals in each category by simply changing the rules in subtle ways from (say): Hospitals receive a three star if they achieve a three on *all* of the three outcome variables to (say): Hospitals receive a three star if they achieve a three on *any* of the three outcome variables.

The following figure shows the new (categorical) composite based on the above decision rules. The new composite (black dots) therefore take values of exactly 0, 1, 2 or 3 only and the uncertainty intervals will equally cover (potentially) the same range.

Figure 23: Uncertainty intervals around a new composite score constructed from thresholds



The correlation with the original composite index is 0.53 and between the two sets of hospital rankings is 0.50 (though once again the composite takes a different form and many of the rankings are also shared since there are fewer permutations with categorical data). The correlation with the previous composite index (to examine the transformation to categorical variables) is 0.53 and between the two sets of hospital rankings is 0.49.

The best ranked hospital from the original composite index now receives a score of 2 whereas the worst ranked hospital receives a score of 0. In the original star rating system these hospitals received a two and one star status respectively. The biggest jump in rankings is for a hospital which originally ranked 112 on the composite (near the bottom) but scored a 2 star status on this composite (a jump of effectively 95 places and an average jump for hospitals of 32 places). Even comparing the changes to the previous composite transformation to categorical variables, the introduction of decision rules leads to one hospital jumping 89 places and an average jump of 22 positions.

The potential for hospitals to change ranking is therefore dramatic when these sorts of decision rules or algorithms are applied to construct a composite index and potentially small and subtle changes to the rules can materially affect the outcome for individual hospitals.

The following table shows the frequency distribution for the number of times that the hospitals are ranked in each of the categories. This is done for a sample of 10 hospitals over the 1000 repetitions. It gives the percentage of times that a hospital appears in each of the categories alongside the category in which it actually appears in the composite produced.

Table 34: Frequency distribution of a sample of 10 hospitals on the new composite index constructed from decision rules

Hospital	New composite category	Percentage of times in simulations that composite is given a score of:			
		0	1	2	3
1	0	**100**	0	0	0
2	0	**82**	18	0	0
3	0	**66**	0	34	0
4	1	2	**61**	0	38
5	1	0	**100**	0	0
6	2	38	2	**32**	28
7	2	19	0	**81**	0
8	2	0	0	**100**	0
9	3	38	2	0	**61**
10	3	0	0	44	**56**

The above table results suggest that the stability of the new composite score constructed from the decision rules is not always very high. Particularly for the better performing hospitals given a rating of three, the number of times they can be counted as achieving composite scores in other categories is quite high, compared to the hospitals attaining a zero rating on the composite. These tend to score zeros and ones more consistently. Poor performance seems to be more stable therefore than good performance as measured in this context.

This is shown in the following two figures where zero rated hospitals often achieve this rating 100 percent of the time in the simulations, whilst for hospitals given a rating of three, shown in the following figure, this is never achieved 100 percent of the time.

Figure 24: Frequency distribution of zero rated hospitals on the new composite index constructed from decision rules

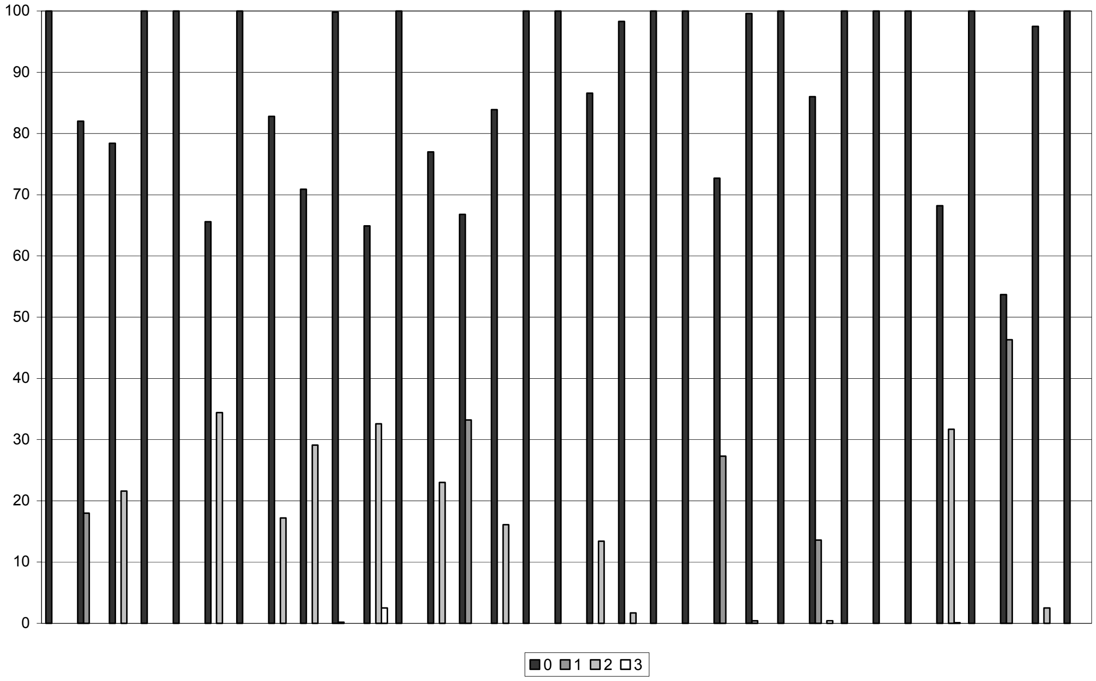
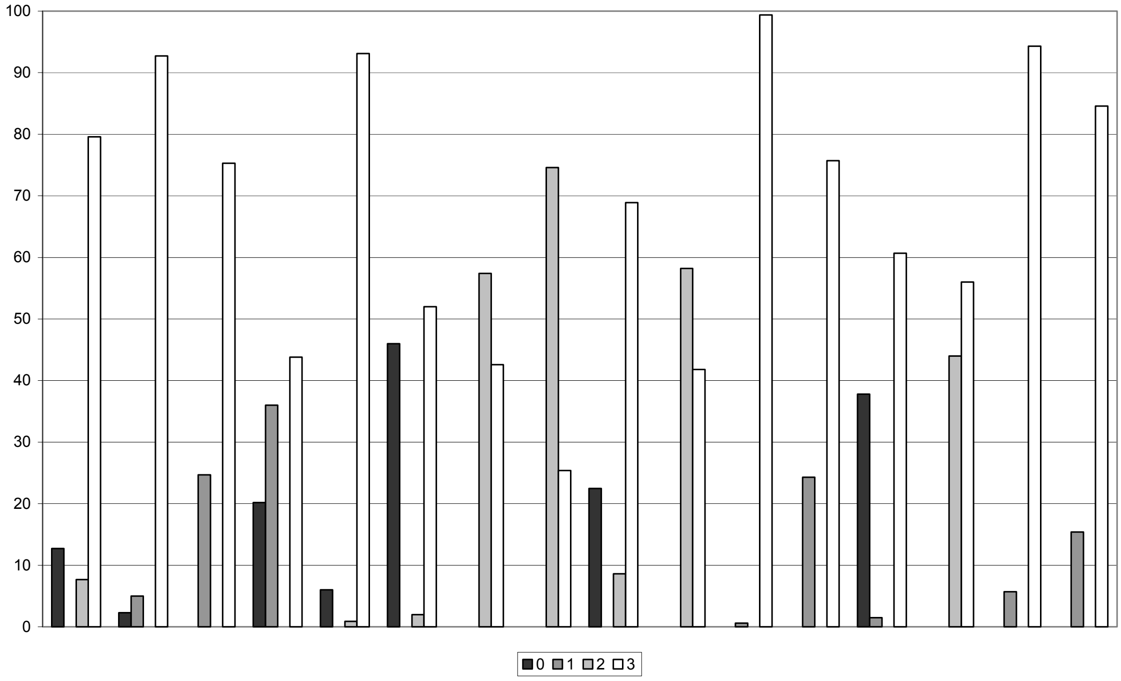


Figure 25: Frequency distribution of hospitals rated three on the new composite index constructed from decision rules



These results seem to agree with the frequency distributions of the rankings of the best (ranked 1) and worst (ranked 117) hospitals on the original composite. These also showed greater stability in the ranking of the worst hospital than the best hospital over the 1000 simulations.

7. Summary and conclusions

Composite indices are a useful communication and political tool for conveying summary performance information in a relatively simple way and signalling policy priorities. They are used widely in health care and other sectors, nationally and internationally. Composite performance indicators have a number of advantages, such as focusing attention on important policy issues, offering a more rounded assessment of performance and presenting the 'big picture' in a way in which the public can understand. It is likely therefore that they will continue to be used in the future in many policy areas.

However, it is important to recognise that the construction of composite indicators is not straightforward and many methodological issues need to be addressed carefully if the results are not to be misinterpreted and manipulated. These issues are more than mere technical pre-occupations. The use and publication of composite performance measures can generate both positive and negative behavioural responses and if significant policy decisions rest on the outcome of the composite (such as Foundation Status for hospitals in the NHS), it is important to have a clear understanding of the potential risks involved in constructing a composite and arriving at such a ranking.

Using data from the NHS Star Ratings to illustrate the methodological issues at each stage of construction of a composite indicator, this report has highlighted the following considerations that deserve careful attention:

- Choice of the units to be assessed. This raises issues such as the degree to which the units chosen are able to influence the performance measures against which they are assessed or whether responsibility falls in part outside the boundary of the unit; and the degree of heterogeneity between units of different types.
- Choice of indicators included in the composite. Effort will be expended on trying to achieve on the included indicators, at the potential expense of excluded indicators which may be as important but simply more difficult to measure. Ultimately composites can only draw on data that is available. The composite can only be as good as the data on which it is based and issues of data reliability, levels of aggregation, and choice of different types of indicator (for example threshold or continuous variables) and their incentive effects have been highlighted.
- Transforming the variables for aggregation. Various options are available (such as rankings, z-scores, logarithms) all with advantages and disadvantages, such as sensitivity to outliers, the loss of interval level information, and sensitivity to weighting. Transformation to a categorical scale is the most widely used method in the UK public sector but can be problematic, since it depends largely on the choice of thresholds which may be selected arbitrarily and can be highly subjective. These different transformations produce important incentive effects and future work could explore in greater detail the impact these choices have on the construction of the composite.
- Choice of a weighting system. The weights applied to each individual indicator prior to aggregation can have a significant impact on the rankings of individual units within the composite. Weights may be derived using statistical methods or they may be chosen to reflect the preferences of specific stakeholders. All too often the choice of weights appears to be ad hoc and arbitrary with a lack of consideration for whose preferences the set of weights reflect and how robust these may be. More work could be done on the sensitivity of weighting structures to the construction of the composite as well as the use of different elicitation methods to extract weights for inclusion in a composite index. Future work could also examine the underpinnings of the relationship between the collinearity of indicators and changes in the weight structure.
- Combining the indicators. Use of decision rules or algorithms about how the scores on individual indicators influence the composite, is a common approach in the public sector. Rather than just adding the indicators, hurdles may be introduced in order to reflect the requirement for attainment of a minimum standard. The nature of the rules will influence the incentives to improve performance, depending on the effort required and the nature of the reward schedule.
- Adjusting for exogenous factors that influence performance. How exogenous influences on measured performance are best taken into account is an important area for future research as well as the relationship between environmental circumstances and the setting of ‘fair’ budgets to reflect that and enable all units an equal opportunity to secure an equal composite score.
- Decomposing variation on performance indicators. The report explores the decomposition of variation on performance indicators (and the composite).

Isolating the element of variation within performance measurement which is truly random and beyond managerial influence is a challenge. There is as yet no established methodology for doing this and the most appropriate ways of separating the different sources of variation in performance is an important avenue for future research.

- Allowing for uncertainty. Given the difficulties associated with measuring performance and the nature of decisions taken at each stage of the process of constructing the composite, the use of sensitivity analysis to explore the robustness of performance rankings is vital.

The empirical analysis using the Star Ratings data demonstrated clearly the importance of each of the above factors, showing that changes in the weights, thresholds and decision rules for combining the indictors materially affects the score and rank correlations of hospitals. The empirical analysis has also highlighted several areas where more work can be done to help analysts understand the impact the technical choices can have on the outcome of the composite.

From a policy point of view, the conceptual limits of composite indicators should be borne in mind and their publication should be accompanied with explanations of the choice of indicators, transformation method, and weighting structure. Explanations of the limits of the composite may help with its interpretation and also with making the process more transparent so that it can be clear what policy objectives are being maximised. This may also make the results more acceptable to participants and may make the reward and penalty schedule attached to the composite, more palatable. Any composite index must be used cautiously and may be supplemented with other types of performance information. The use of ‘soft’ data in the healthcare sector may be very important where several aspects of performance cannot be captured by hard numbers or league tables (Goddard *et al*, 1999).

Notwithstanding the importance of considering these important methodological and policy issues, some pragmatism in the approach to composites may be appropriate. Often the construction of composites that are less than ideal, may nevertheless lead to important empirical and policy analyses, as has been the case with both the WHO country rankings of health care systems and the star ratings. Their publication and release may lead to the search for better analytical methods and improvements in data capture.

Technical and analytical issues in the design of composite indicators clearly have important policy implications. This report highlights the issues which need to be taken into account in the construction of robust composite indicators so that they can be designed in ways which will minimise the potential for producing misleading performance information. If such issues are not addressed, composite measures may fail to deliver the expected improvements in performance or may even induce unwanted side-effects.

8. References

Abayomi, K., Gelman, A. & Srebotnjak, T. (2002) *Imputing missing values*, Appendix 3 to World Economic Forum's 2002 environmental sustainability report, Center for International Earth Science Information Network (CIESIN): Columbia University. http://www.ciesin.org/indicators/ESI/

Adriaanse, A. (1993) *Environmental policy performance indicators: A study on the development of indicators for environmental policy in the Netherlands*, SDV Publishers: The Hague.

Almeida, C., Braveman, P., Gold, M.R., Szwarcwald, C.L., Ribeiro, J.M., Miglionico, A., Millar, J.S., Porto, S., Costa, N.R., Rubio, V.O., Segall, M., Starfield, B., Travessos, C., Uga, A., Valente, J. & Viacava, F. (2001) Methodological concerns and recommendations on policy consequences of the World Health Report 2000, *The Lancet*, 357(9269): 1692-7.

Appleby, J. & Mulligan, J. (2000) *How well is the NHS performing? A composite performance indicator based on public consultation*, King's Fund: London.

Appleby, J. & Street, A. (2001) Health system goals: life, death and … football, *Journal of Health Services Research*, 6(4): 220-225.

Audit Commission (2003a) *Technical manual on the use of performance indicators in CPA*, Audit Commission: London. http://www.audit-commission.gov.uk/cpa/downloads/TechnicalManualUsePIsCPA.pdf

Audit Commission (2003b) *Technical manual on the use of performance indicators in CPA: update for 2003/04 and 2004/05*, Audit Commission: London. http://www.audit-commission.gov.uk/cpa/downloads/CPASTCCTechnicalManualonPIs.doc

Audit Commission (2003c) *Comprehensive Performance Assessment Guidance for district councils Version F1.1*, Audit Commission: London. http://www.audit-commission.gov.uk/cpa/downloads/april1.pdf

Benefit Fraud Inspectorate (2003) *CPAs of single tier Authorities*, BFI: Harrogate. http://www.bfi.gov.uk/about/products/cpa/local_authority.asp

Berki, S.E. (1972) *Hospital economics*, Studies in Social and Economic Process, Lexington Books, D.C. Heath and Company: Massachusetts.

Bernstein, A. & Gauthier, A. (1999) Choices in health care: what are they and what are they worth? *Medical Care Research and Review*, 56(supplement 1): 5-23.

Butler, J.R.G. (1995) *Hospital cost analysis*, Kluwer Academic Publishers: Dordrecht.

Canadian Institute for Health Information (2001a) *Health Care in Canada 2001: A Second Annual Report*, Canadian Institute for Health Information: Ottawa.

Canadian Institute for Health Information (2001b) *Health Indicators 2001*, Canadian Institute for Health Information: Ottawa.

Commission for Health Improvement (2003a) *NHS performance ratings acute trusts, specialist trusts, ambulance trusts 2002/2003*, Commission for Health Improvement: London. http://www.chi.nhs.uk/ratings/

Commission for Health Improvement (2003b) *Rating the NHS: A question and answer guide to the Finsbury Rules, How CHI CGR scores will affect 2003 star ratings*, Commission for Health Improvement: London. http://www.chi.nhs.uk/eng/ratings/finsbury_rules.shtml

Commission for Social Care Inspection (2003) *The Commission for Social Care Inspection*, Commission for Social Care Inspection: London. http://www.doh.gov.uk/csci/

Crombie, I.K. & Davies, H.T.O. (1998) Beyond health outcomes: The advantages of measuring process, *Journal of Evaluation in Clinical Practice*, 4(1): 31-38.

Cutler, T. (2002) Star or black hole? *Community Care*, 30 May 2002, 40-41.

D'Agostino, R.B., Balanger, A. & D'Agostino, R.B.Jr. (1990) A suggestion for using powerful and informative tests for normality, *The American Statistician*, 44(4): 316-321.

Davies, H.T.O. & Crombie, I.K. (1995) Assessing the quality of care, *British Medical Journal*, 311: 766.

Davies, H.T.O., Dawson, D. & Smith, P.C. (2002) *Should process or outcome measures be used for performance management?*, in Smith, P.C. (Ed.), Some principles of performance measurement and performance improvement, Report prepared for the Commission for Health Improvement: University of York, February 2002.

Department of Health (2000) *Quality and Performance in the NHS Performance Indicators: July 2000*, Department of Health: London. http://www.doh.gov.uk/nhsperformanceindicators/indicators2000.htm.

Department of Health (2001) *NHS Performance Ratings: Acute Trusts 2000/01*, Department of Health: London. http://www.doh.gov.uk/performanceratings/2001/index.html

Department of Health (2002a) *NHS Performance Ratings and Indicators: Acute Trusts, Specialist Trusts, Ambulance Trusts, Mental Health Trusts 2001/02*, Department of Health: London. http://www.doh.gov.uk/performanceratings/2002/index.html

Department of Health (2002b) *Performance Ratings Methodology - Acute NHS Hospital Trusts*, Department of Health: London. http://www.doh.gov.uk/performanceratings/2002/method_acute.html

Department of Health (2002c) *A guide to social services performance "star" ratings*, Social Services Inspectorate: London. http://www.doh.gov.uk/pssratings/wholedocument.pdf

Department of Health (2003) *How SSI assess performance*, Social Services Inspectorate: London. http://www.doh.gov.uk/ssi/performance.htm

DETR (2000) *Indices of deprivation*, Regeneration Research Summary Number 31 Department of the Environment, Transport and the Regions, HMSO: London. http://www.odpm.gov.uk/stellent/groups/odpm_urbanpolicy/documents/downloadable/odpm_urbpol_021680.pdf

DLTR (2001) *Strong Local Leadership – Quality Public Services*, CM5237, Department for Transport, Local Government and the Regions, HMSO: London. http://www.europartnership.com/Documents/wp_part1.pdf

Dolan, P., Gudex, C., Kind, P. & Williams, A. (1996) Valuing health states: A comparison of methods, *Journal of Health Economics*, 15: 209-231.

Donabedian, A. (1966) Evaluating the quality of medical care, *Milbank Memorial Fund Quarterly*, XLIV(3): 166-206.

Dranove, D., Kessler, D., McClellan, M. & Satterthwaite, M. (2002) *Is more information better? The effects of report cards on health care providers*, National Bureau for Economic Research Working Paper 8697, NBER: Cambridge, Massachusetts.

Ellaway, A. (1997) Are single indicators of deprivation as useful as composite indicators in predicting morbidity and mortality: Results from the Central Clyneside Conurbation, *Health Bulletin*, 55(5): 283-4.

European Commission (2000) *Business Climate Indicator*, DG ECFIN, European Commission: Brussels.

European Commission (2001a) *Summary Innovation Index*, DG Enterprise, European Commission: Brussels.

European Commission (2001b) *Internal Market Scoreboard*, DG MARKT, European Commission: Brussels.

Eurostat (2000) *Index of Environmental Friendliness*, Statistics Finland: Helsinki. http://www.stat.fi/tk/yr/ye22_en.html

Fagerberg, J. (2001) *Europe at the crossroads: The challenge from innovation-based growth*, in Lundvall, B. & Archibugi, D. (Eds.) The Globalising Learning Economy, Oxford University Press: Oxford.

Freudenberg, M. (2003) *Composite indicators of country performance: A critical assessment*, OECD STI Working paper DSTI/DOC 2003/16, OECD: Paris.

Fried, H.O., Schmidt, S.S. & Yaisawarng, S. (1999) Incorporating the operating environment into a nonparametric measure of technical efficiency, *Journal of Productivity Analysis*, 12(3): 249-67.

Goddard, M., Mannion, R. & Smith, P.C. (1999) Assessing the performance of NHS hospital Trusts: The role of 'hard' and 'soft' information, *Health Policy*, 48: 119-34.

Goddard, M. & Smith, P.C. (2001) Performance measurement in the new NHS, *Health Policy Matters*, 3: 1-8.

Gravelle, H., Jacobs, R., Jones, A.M. & Street, A. (2004) Comparing the efficiency of national health systems: a sensitivity analysis of the WHO approach, *Applied Health Economics and Health Policy,* forthcoming.

Gravelle, H. & Smith, P.C. (2002) *Sources of variations in measured performance*, in Smith, P.C. (Ed.), Some principles of performance measurement and performance improvement, Report prepared for the Commission for Health Improvement: University of York, February 2002.

Greene, W. (2003) *Distinguishing Between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems*, Department of Economics, Stern School of Business: New York University.

Hannan, E.L., Kilburn, H., Racz, M., Shields, E. & Chassin, M.R.(1994) Improving the outcomes of coronary artery bypass surgery in New York State, *Journal of the American Medical Association*, 271: 761-6.

Hauck, K., Shaw, R. & Smith, P. (2002), Reducing avoidable inequalities in health: a new criterion for setting health care capitation payments, *Health Economics*, 11(8): 667-677.

Hauck, K., Rice, N. & Smith, P. (2002), The influence of health care organisations on health system performance, *Journal of Health Services Research and Policy*, 8(2): 68-74.

HEFCE (2003) *Research assessment*, The Higher Education Funding Council for England (HEFCE): Bristol. http://www.hefce.ac.uk/research/assessment/

HERO (2001) *A Guide to the 2001 Research Assessment Exercise*, Higher Education & Research Opportunities in the United Kingdom, HEFCE: Bristol. http://www.hero.ac.uk/rae/Pubs/other/raeguide.pdf

Hofstetter, P., Braunschweig, A., Mettier, M., Müller-Wenk, R. & Tietje, O. (1999) Dominance Analysis in the Mixing Triangle, A Graphical Decision Support Tool for Comparisons with LCA, *Journal of Industrial Ecology*, 3(4): 97-115.

Home Office (2002) *Police Performance Monitoring 2001/02*, Home Office, London. http://www.policereform.gov.uk/docs/wholeperformancemonitorsdoc.pdf

Iezzoni, L.I., Schwartz, M., Ash, A.S., Hughes, J.S., Daley, J. & Mackiernan, Y.D. (1995) Using severity-adjusted stroke mortality rates to judge hospitals, *International Journal for Quality in Health Care*, 7(2): 81-94.

Iezzoni, L.I., Schwartz, M., Ash, A.S., Hughes, J.S., Daley, J. & Mackiernan, Y.D. (1996) Severity measurement methods and judging hospital death rates for pneumonia, *Medical Care*, 34(1): 11-28.

Jacobs, R. (2001) Alternative methods to examine hospital efficiency: Data Envelopment Analysis and Stochastic Frontier Analysis, *Health Care Management Science*, 4(2): 103-16.

Jacobs, R. & Dawson, D. (2002) Variation in Trust unit costs in the NHS, *Journal of Health Services Research and Policy*, submitted.

Jacobs, R. & Smith, P. (2003) *A descriptive analysis of general acute trust star ratings*, Centre for Health Economics: University of York.

Jencks, S.F., Cuerdon, T., Burwen, D.R., Fleming, B., Houck, P.M., Kussmaul, A.E., Nilasena, D.S., Ordin, D.L. & Arday, D.R. (2000) Quality of medical care delivered to Medicare beneficiaries: A profile at state and national levels, *Journal of the American Medical Association*, 284(13): 1670-6.

Jencks, S.F., Huff, E.D. & Cuerdon, T. (2003) Change in the quality of care delivered to Medicare beneficiaries, 1998-1999 to 2000-2001, *Journal of the American Medical Association*, 289(3): 305-12.

Jensen, U. (2000) Is it efficient to analyse efficiency rankings? *Empirical Economics*, 25: 189-208.

Joint Research Centre (2002) *State of the art report on current methodologies and practices for composite indicator development*, Report prepared by the Applied Statistics Group, Institute for the Protection and Security of the Citizen: European Commission, June 2002.

Kmietowicz, Z. (2003) Star rating system fails to reduce variation, *British Medical Journal*, 327: 184.

Little, R. & Rubin, D. (1987) *Statistical analysis with missing data*, Wiley: New York.

Lubalin, J.S., & Harris-Kojetin, L.D. (1999) What do consumers want and need to know in making health care choices?, *Medical Care Research and Review*, 56(Supplement 1): 67–102.

Mannion, R. & Goddard, M. (2001) Impact of published clinical outcomes data: Case study in NHS hospital trusts, *British Medical Journal*, 323: 260-3.

Marcenes, W.S. & Sheiham, A. (1993) Composite indicators of dental health: Functioning teeth and the number of sound-equivalent teeth (T-Health), *Community Dentistry and Oral Epidemiology*, 21: 374-378

Marshall, M.N., Hiscock, J. & Sibbald, B. (2002) Attitudes to the public release of comparative information on the quality of general practice: qualitative study, *British Medical Journal*, 325: 1277-81.

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(14): 1866-74.

McKee, M. (2001) Measuring the efficiency of health systems, *British Medical Journal*, 323: 295-296.

McLaughlin, C.G. (1999) Health care consumers: choices and constraints, *Medical Care Research and Review*, 56(supplement1): 24-59.

Miller, N. (2002) Missing the target, *Community Care*, 38.

Moldan, B., Billharz, S. & Matravers, R. (1997) *SCOPE 58 Sustainability indicators: A Report on the Project on Indicators of Sustainable Development*, Scientific Committee on Problems of the Environment, John Wiley Publishers: New York.

Mooney, C.Z. (1997) *Monte Carlo simulation*, A Sage University Paper, No. 116 in Series: Quantitative Applications in the Social Sciences, Sage Publications Ltd: London.

Muldur, U. (2001) *Technical annex on structural indicators: Two composite indicators to assess the progress of Member States in their transition towards a knowledge-based economy*, DG RTD, European Commission: Brussels.

Mullen, P. & Spurgeon, P. (2000) *Priority setting and the public*, Radcliffe Medical Press: Abingdon.

Navarro, V. (2000) Assessment of the World Health Report, *The Lancet*, 356(9241): 1598–1601.

Navarro, V. (2001) World Health Report 2000: response to Murray and Frenk, *Lancet*, 357(9269): 1701-2.

Navarro, V. (2002) The World Health Report 2000: can health care systems be compared using a single measure of performance? *American Journal of Public Health*, 92(1): 31-34.

Nord, E. (2002) Measures of goal attainment and performance: a brief, critical consumer guide, *Health Policy*, 59(3): 183-91.

Nutley, S. & Smith, P.C. (1998) League ables for performance improvement in health care, *Journal of Health Services Research and Policy*, 3(1): 50-7.

ODPM (2003) *Local government performance, Best Value Performance Indicators*, HMSO: London. http://www.bvpi.gov.uk/pages/index.asp

Office for Standards in Education (2002a) *The development of the Comprehensive Performance Assessment framework for the education sector: Briefing paper for Chief Education Officers*, Ofsted: London. http://www.ofsted.gov.uk/publications/docs/3021.pdf

Office for Standards in Education (2002b) *York comprehensive performance assessments*, Ofsted: London. http://www.ofsted.gov.uk/reports/manreports/642.pdf

Office for Standards in Education (2003) *Inspecting schools: Framework for inspecting schools (Effective from September 2003)*, Ofsted: London. http://www.ofsted.gov.uk/publications/docs/3266.pdf

OFSTED and DfES (2002) *Guidelines for the Education Profile of CPA: The contents and operation of the education profile for the 2002 Comprehensive Performance Assessment*, Office for Standards in Education: London. http://www.ofsted.gov.uk/publications/docs/3020.doc

Page, S. & Cramer, K. (2001) Maclean's Rankings of Health Care Indices in Canadian Communities, 2000: Comparisons and Statistical Contrivance, *Canadian Journal of Public Health*, 92(4): 295-98.

Porter, M. & Stern, S. (1999) *The new challenge to America's prosperity: Findings from the Innovation Index*, Council of Competitiveness, Washington D.C. http://www.compete.org/pdf/innovation.pdf

PRé (Product Ecology Consultants) (2000) *Eco- Indicator 99 Principles*, Product Ecology Consultants: Amersfoort, The Netherlands. http://www.pre.nl/eco-indicator99/eco-indicator_99.htm

Ryan, M. & Farrar, S. (2000) Using conjoint analysis to elicit preferences for health care, *British Medical Journal*, 320: 1530-3.

Saaty, R.W. (1987) The analytic hierarchy process: what it is and how it is used, *Mathematical Modelling*, 9: 161-76.

Schneider, E.C. & Epstein, A.M. (1996) Influence of cardiac surgery performance reports on referral practices and access to care, *New England Journal of Medicine*, 335: 251-6.

Schneider, E.C. & Lieberman, T. (2001) Publicly disclosed information about the quality of health care: response of the US public, *Quality in Health Care*, 10: 96-103.

Shaw, R., Dolan, P., Tsuchiya, A., Williams, A., Smith, P.& Burrows, R. (2001) *Development of a questionnaire to elicit preferences regarding health inequalities*, Occasional Paper Number 40, Centre for Health Economics, University of York: York.

Sheldon, T., Maynard, A. & Watt, I. (2001) Promoting quality in the NHS, *Health Policy Matters*, 4: 1-4.

Simar, L. & Wilson, P.W. (2002) *Estimation and inference in two-stage semi-parametric models of production processes*, Louvain: Institut de Statistique, Universite Catholique de Louvain.

Smith, P. (1995) On the Unintended Consequences of Publishing Performance Data in the Public Sector, *International Journal of Public Administration*, 18: 277-310.

Smith, P. (2002) *Developing composite indicators for assessing health system efficiency*, in Smith, P.C. (ed.) Measuring up: Improving the performance of health systems in OECD countries, OECD: Paris.

Smith, P. (2003) Formula funding of public services: an economic analysis, *Oxford Review of Economic Policy*, 19(2): 301-322.

Smith, P., Rice, N. & Carr-Hill, R. (2001) Capitation funding in the public sector, *Journal of the Royal Statistical Society*, Series A, 164(2): 217-57.

Snelling, I. (2003) Do star ratings really reflect hospital performance? *Journal of Health Organization and Management*, 17: 210-223.

SPRG (2001) *Report of the Scientific Peer Review Group on Health Systems Performance Assessment*, Scientific Peer Review Group (SPRG), WHO: Geneva. http://www.who.int/health-systems-performance/sprg/report_of_sprg_on_hspa.htm

Storrie, D. & Bjurek, H. (2000) *Benchmarking European labour market performance with efficiency frontier techniques*, Discussion paper FS I 00-211, Department of Social Work: Göteborg University, Sweden.

Street, A. (2002) The resurrection of hospital mortality statistics in England, *Journal of Health Services Research and Policy*, 7(2): 104-10.

Street, A. (2003) How much confidence should we place in efficiency estimates?, *Health Economics*, 12(11): 895-907.

The Quality Assurance Agency for Higher Education (2003) The Quality Assurance *Agency for Higher Education: an introduction*, QAA: Gloucester. http://www.qaa.ac.uk/aboutqaa/qaaintro/intro.htm#3

United Nations (2003) *Human Development Report 2003, Millennium Development Goals: A compact among nations to end human poverty*, United Nations Development Programme, Oxford University press: Oxford. http://www.undp.org/hdr2003/

Williams, A. (2001) Science or marketing at WHO?, *Health Economics*, 10: 93-100.

World Economic Forum (2002) *Environmental Sustainability Index*, Center for International Earth Science Information Network (CIESIN): Columbia University. http://www.ciesin.org/indicators/ESI/

World Health Organisation (2000) *The World Health Report 2000, Health systems: improving performance*, WHO: Geneva.

9. Appendix

Table 35: Variable names and descriptions

Performance indicator	Indicator variable and other underlying data provided	Variable name
Key targets		
Eighteen month inpatient waits	Number of patients reported across the year as waiting more than 18 months for inpatient treatment	inpwt18mn
Fifteen month inpatient waits	Number of patients waiting more than 15 months for an inpatient admission	wt15mn
Twenty six week outpatient waits	Number of patients waiting more than 6 months (26 weeks) for an outpatient appointment	outwt26wk
Twelve hour trolley waits	Number of patients waiting more than 12 hours for admission via A&E	a_e12hrwt
Cancelled operations	Percentage of elective admissions cancelled at the last minute for non-clinical reasons	cancelopspc
	Number of last minute cancellations	cancelops
Two week cancer waits	Percentage of patients seen within two weeks of urgent GP referral for suspected cancer to outpatient appointment with specialist	cancerwt2wkpc
Improving working lives	Progress towards achievement of Improving Working Lives (IWL) standard 'practice status'	iwl
Hospital cleanliness	Whole hospital score for cleanliness, formulated against Patient Environment Action Team (PEAT) visits (Values range from 1 to 4)	cleanliness
Financial management	Achievement of financial balance without unplanned financial support	finbalpc
	Bottom Line Month 12 (£000's)	bottomline
	Forecast Turnover Month 12 (£000s)	turnover
	Unplanned Financial Support (£000s)	fin_support
Clinical focus		
Clinical negligence	Level of compliance against Clinical Negligence Scheme for Trusts (CNST) risk management standards (Values range from 0 to 3)	cnst
Death within 30 days of surgery (non-elective admissions)	Deaths within 30 days of surgery for non-elective admissions to hospital, per 100,000 patients (age and sex standardised, includes deaths in hospital and after discharge)	d_esurgstd
Death within 30 days of a heart bypass operation	Deaths within 30 days of a Coronary Artery Bypass Graft (CABG), per 100,000 patients (age, sex and method of admission standardised, includes deaths in hospital and after discharge)	d_heartpl
Emergency readmission to hospital following discharge	Emergency readmissions to hospital within 28 days of discharge (all ages), as a percentage of live discharges (age and sex standardised)	readmisnpc
	Number of readmissions	readmisnstd
Emergency readmission to hospital following discharge for children	Emergency readmissions of children to hospital within 7 days of discharge following medical treatment (ages 0-15), as a percentage of live discharges (age and sex standardised)	readm_child
Emergency readmission to hospital following treatment for a fractured hip	Emergency readmissions to hospital within 28 days of discharge following treatment for a fractured hip, as a percentage of live hip fracture discharges (age and sex standardised)	readm_hip
Emergency readmission to hospital following treatment for a stroke	Emergency readmissions to hospital within 28 days of discharge following a stroke, as a percentage of live stroke discharges (age and sex standardised)	readm_stroke
Returning home following hospital treatment for fractured hip	Percentage of patients discharged back to usual place of residence within 28 days of emergency admission to hospital with a hip fracture, all ages (age and sex standardised)	dis_hippc
Returning home following hospital treatment for stroke	Percentage of patients discharged back to usual place of residence within 56 days of emergency admission to hospital with a stroke, all ages (age and sex standardised)	dis_strokepc
Patient focus		
Six month inpatient waits	Percentage of patients waiting less than 6 months for an inpatient admission	wait6pc
	Number of patients waiting less than 6 months for inpatient admission	wait6mn
Total inpatient waits	Total number of patients waiting for an inpatient appointment (% of planned target achieved)	wttargetpc
Thirteen week outpatient waits	Percentage of patients seen within 13 weeks of GP written referral for first outpatient appointment	outwt13wkpc
Total time in A&E	Percent of total A&E attendances spending less than 4 hours in A&E from arrival to admission, transfer or discharge	a_e4hrwtpc
	Attendances spending less than 4hrs in A&E	a_e4hrwt
Cancelled operations not admitted within a month	Percentage of elective admission patients not admitted within one month of last minute cancellation	cancelop1mnpc
	No of patients not admitted within 1 month of cancelled operation	cancelop1mn
Heart operation waits	Number of patients waiting more than 12 months for a coronary artery bypass graft (CABG) and percutaneous transluminal coronary angioplasty (PTCA)	wt_heart

	graft (CABG) and percutaneous transluminal coronary angioplasty (PTCA)	
Breast cancer treatment	Percent treated within 2 weeks from diagnosis to treatment for patients newly diagnosed with breast cancer	breastwt2wkpc
Delayed discharges	Percentage of patients whose discharge from hospital was delayed	delay_dispc
	No of delayed discharges from acute beds	delay_dis
Inpatient survey - coordination of care	Combined score of questions around organisation of emergency care, organisation of admissions process, staff giving conflicting information and members of staff taking overall charge of care	inp_survey_coord
Inpatient survey - environment and facilities	Combined score of questions around cleanliness and quality and amount of food provided	inp_survey_env
Inpatient survey - information and education	Combined score of questions around staff's responses to questions, explanations of medication and information about possible problems following transfer of care	inp_survey_inf
Inpatient survey - physical and emotional needs	Combined score of questions around noise, adequate control of pain, assistance during mealtimes and discussions with staff about anxieties	inp_survey_phys
Inpatient survey - prompt access	Combined score of questions around length of waiting list, notice given of admission, change to admission date and waiting time to get to a ward through admission or A&E	inp_survey_acc
Inpatient survey - respect and dignity	Combined score of questions around privacy for discussion, examination and treatment, mixed sex facilities and whether patient was treated with dignity and respect	inp_survey_resp
Capacity and capability focus		
Data quality	Summary measure of Hospital Episode Statistics (HES) data quality for NHS trusts with in-patient activity	dqi_pc
Staff opinion survey	Responses from NHS-employed staff opinion survey on satisfaction with employer	staff_survey
Junior doctors' hours	Percentage of Junior Doctors complying in full with the New Deal on Junior Doctors' Hours	jundocpc
	Number of junior doctors complying with New Deal	jundoc_cmpl
	Number of junior doctors in post	jundoc_post
Sickness absence rate	Amount of time lost through absences as a percentage of staff time available for directly employed NHS staff	sick_rate
Information governance	Achievement of information governance targets (Values range from 0 to 36)	info_gov
CHI review	Clinical governance review covering areas of risk management, clinical audit, research and education, patient involvement, information management, staff involvement and education, training and development	chi_review

Table 36: Descriptive statistics for raw data

variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
inpwt18mn	181	4.608	0	42.188	0	537	11.618	143.466	0.000
wt15mn	181	1.238	0	16.575	0	223	13.341	178.998	0.000
outwt26wk	181	4.508	0	36.437	0	352	9.083	84.966	0.000
a_e12hrwt	160	9.663	0	43.224	0	408	6.935	56.273	0.000
cancelops	184	108.234	78.0	97.778	1	506	1.705	6.374	0.000
cancelopspc	184	1.604	1.104	1.436	0.032	11.409	2.518	14.355	0.000
cancerwt2wkpc	168	93.972	96.109	6.194	72.903	100	-1.321	4.140	0.000
iwl	181	1	1	0	1	1	.	.	.
cleanliness	171	3.392	3	0.490	3	4	0.443	1.196	.
bottomline	181	-133.227	7	1386.385	-11487	6762	-3.334	36.340	0.000
turnover	181	137439	112763	82659	10682	550749	1.744	7.215	0.000
finbalpc	181	-0.131	0.007	0.885	-6.620	2.164	-3.563	23.824	0.000
fin_support	181	278.343	0	1486.878	0	12420	6.436	46.036	0.000
cnst	171	1.053	1	0.512	0	3	0.350	4.708	0.004
d_esurgstd	149	2951.28	2935.15	528.296	1438.72	4386.48	-0.175	3.051	0.606
d_heartpl	27	2872.86	2893.54	710.605	1342.70	4591.11	0.243	3.027	0.694
readmisnstd	150	3182.29	2628	1792.214	924	11158	1.683	7.041	0.000
readmisnpc	150	6.033	5.959	0.909	4.186	8.955	0.513	3.130	0.039
readm_child	146	5.121	4.818	1.964	1.863	20.765	3.582	29.053	0.000
readm_hip	149	8.036	7.753	2.383	1.774	14.774	0.429	3.181	0.077
readm_stroke	149	7.322	6.971	2.392	2.752	18.069	0.895	5.804	0.000
dis_hippc	148	47.925	46.824	8.315	28.307	75.074	0.352	3.104	0.173
dis_strokepc	149	50.488	50.234	5.336	35.020	67.353	-0.080	3.321	0.542
wait6pc	180	78.830	78.574	8.730	60.830	100	0.319	2.723	0.163
wait6mn	180	4261.95	3796	2456.301	37	12490	0.909	3.776	0.000
wttargetpc	181	3.104	-0.101	16.961	-25.75	179.012	6.502	65.976	0.000
outwt13wkpc	175	74.647	74.613	9.152	49.887	97.677	-0.150	2.786	0.641
a_e4hrwtpc	163	77.123	79.195	13.802	29.914	100	-0.604	3.004	0.014
a_e4hrwt	163	12762.75	11777	5705.009	1664	32750	1.285	4.790	0.000
cancelop1m~c	184	0.371	0.132	0.696	0	7.209	5.830	53.034	0.000
cancelop1mn	184	25.554	10	43.881	0	315	3.447	18.579	0.000
wt_heart	32	0.625	0	3.536	0	20	5.388	30.032	0.000
breastwt2wkpc	147	94.894	99.573	12.533	20.588	100	-3.948	19.882	0.000
delay_dispc	170	4.806	4.087	3.089	0	17.405	0.897	4.160	0.000
delay_dis	170	27.362	23.462	20.296	0	108.308	1.046	4.285	0.000
inp_survey_coord	171	68.007	67.78	4.234	56.71	79.71	0.217	3.328	0.273
inp_survey_env	171	72.736	73.41	5.544	57.90	84.54	-0.531	2.744	0.024
inp_survey_inf	171	68.344	68.18	4.346	56.56	80.67	0.293	3.382	0.136
inp_survey_phys	171	71.087	70.65	4.113	60.56	84.36	0.395	3.829	0.021
inp_survey_acc	171	79.320	79.85	7.678	54.42	93.61	-0.519	2.906	0.030
inp_survey_resp	171	82.309	82.56	4.722	67.93	93.46	-0.334	3.066	0.173
dqi_pc	180	90.961	93.438	7.328	66.825	99.214	-1.504	4.629	0.000
staff_survey	168	3.190	3.187	0.145	2.653	3.684	-0.221	4.387	0.018
jundocpc	171	59.368	59.8	21.160	0	100	-0.254	2.888	0.375
jundoc_cmpl	171	86.202	68	68.581	0	380	1.997	7.990	0.000
jundoc_post	171	149.108	116	111.498	7	738	2.115	8.874	0.000
sick_rate	168	4.564	4.5	0.744	2.5	6.7	0.115	3.149	0.661
info_gov	181	21.785	22	6.083	0	36	-0.588	4.747	0.000

Table 37: Thresholds for scoring indicators

Rating given	Significantly under achieved	Under achieved	Achieved		
	(6)	(–)	(4)		
Score given	1	2	3		
Variable					
inpwt18mn_s	>2	0-2	=0		
wt15mn_s	>10	1-10	<1		
outwt26wk_s	>50	2-50	<2		
a_e12hrwt_s	>75	10-75	<10		
cancelops_s	>5%	1%-5%	<1%		
cancerwt2wk_s	<80%	80%-95%	>95%		
iwl_s			=1		
cleanliness_s			>=3		
finman_s	<-3	-3--1	>-1		
Rating given	Significantly below average	Below average	Average	Above average	Significantly above average
Performance supposed to fall in percentile	1-10	11-30	31-70	71-90	91-100
Score given	1	2	3	4	5
Variable					
cnst_s	=0			=1	2-3
d_esurgstd_s	No confidence interval overlap - higher values		Overlap confidence interval (2889-2956)		No confidence interval overlap - lower values
d_heart_s	No confidence interval overlap - higher values		Overlap confidence interval (2796-3053)		No confidence interval overlap - lower values
readmisn_s	No confidence interval overlap - higher values		Overlap confidence interval (6.10-6.13)		No confidence interval overlap - lower values
readm_child_s	No confidence interval overlap - higher values		Overlap confidence interval (5.40-5.52)		No confidence interval overlap - lower values
readm_hip_s	No confidence interval overlap - higher values		Overlap confidence interval (7.71-8.25)		No confidence interval overlap - lower values
readm_stroke_s	No confidence interval overlap - higher values		Overlap confidence interval (7.14-7.65)		No confidence interval overlap - lower values
dis_hip_s	No confidence interval overlap - lower values		Overlap confidence interval (47.01-48.26)		No confidence interval overlap - higher values
dis_stroke_s	No confidence interval overlap - lower values		Overlap confidence interval (50.24-51.41)		No confidence interval overlap - higher values
wait6mn_s	<67.9%	67.9%-73.5%	73.5%-82.7%	82.7%-90.8%	>90.8%
wttarget_s	>17.6%	4.6%-17.6%	-1.4%- 4.6%	-9.3%- -1.4%	<-9.3%
outwt13wk_s	<63.5%	63.5%-69.6%	69.6%-79.9%	79.9%-86.8%	>86.8%
a_e4hrwt_s	<58.1%	58.1%-70.2%	70.2%-86.1%	86.1%-93.8%	>93.8%
cancelop1mn_s	>0.9%	0.4%-0.9%	0%-0.4%		=0%
wt_heart_s	>0				=0
breastwt2wk_s	<84.5%	84.5%-95.9%	95.9%-100%		=100%
delay_dis_s	>9%	6.1%-9.0%	3.2%-6.1%	1.2%-3.2%	<1.2%
inp_survey_coord_s	<63	63-66.2	66.2-69.8	69.8-73.3	>73.3
inp_survey_env_s	<65.6	65.6-71	71-76.4	76.4-79.4	>79.4
inp_survey_inf_s	<63.5	63.5-66.4	66.4-70.3	70.3-73.8	>73.8
inp_survey_phys_s	<66	66-69.2	69.2-73	73-75.7	>75.7
inp_survey_acc_s	<68.3	68.3-76	76-84.5	84.5-88	>88
inp_survey_resp_s	<76.3	76.3-80.2	80.2-85.2	85.2-88	>88

dqi_s	<79.9%	79.9%-89.7%	89.7%-95.8%	95.8%-97.7%	>97.7%
staff_survey_s	<3	3-3.1	3.1-3.3	3.3-3.4	>3.4
jundoc_s	<31.6%	31.6%-50.8%	50.8%-69.9%	69.9%-84.9%	>84.9%
sick_rate_s	>5.5%	4.9%-5.5%	4.2%-4.9%	3.7%-4.2%	<3.7%
info_gov_s	<15	15-19	19-25	25-30	>30
Rating given	Significant area of weakness (6)	Some strengths (–)	Many strengths (4)	Significant strengths (44)	
Score given	1	2	3	4	
Variable					
chi_review	1	2	3	4	

Table 38: Descriptive statistics for transformed variables

variable	n	mean	median	std.dev	min	max	skewness	kurtosis	Prob>chi2
inpwt18mn_s	186	2.871	3	0.434	1	3	-3.445	13.982	0.000
wt15mn_s	186	2.989	3	0.147	1	3	-13.528	184.005	0.000
outwt26wk_s	186	2.962	3	0.241	1	3	-6.892	52.008	0.000
a_e12hrwt_s	185	2.768	3	0.483	1	3	-1.962	6.071	0.000
cancelops_s	186	2.403	2	0.534	1	3	-0.036	1.888	0.000
cancerwt2wk_s	173	2.532	3	0.576	1	3	-0.765	2.584	0.001
iwl_s	186	3	3	0	3	3	.	.	.
cleanliness_s	186	3	3	0	3	3	.	.	.
finman_s	186	2.704	3	0.652	1	3	-1.959	5.208	0.000
cnst_s	186	4.038	4	0.745	1	5	-2.575	12.472	0.000
d_esurgstd_s	166	2.783	3	1.226	1	5	0.063	2.615	0.548
d_heart_s	29	2.931	3	0.998	1	5	-0.082	4.131	0.217
readmisn_s	166	3.120	3	1.812	1	5	-0.119	1.236	.
readm_child_s	160	3.200	3	1.541	1	5	-0.172	1.711	0.000
readm_hip_s	165	2.867	3	0.630	1	5	-1.507	9.048	0.000
readm_stroke_s	164	2.902	3	0.619	1	5	-1.183	9.870	0.000
dis_hip_s	163	2.926	3	1.086	1	5	-0.028	3.392	0.486
dis_stroke_s	164	2.854	3	0.785	1	5	-0.653	6.057	0.000
wait6mn_s	185	3.011	3	1.098	1	5	0.003	2.500	0.253
wttarget_s	186	3.032	3	1.095	1	5	0.010	2.511	0.275
outwt13wk_s	180	3	3	1.099	1	5	0	2.5	0.265
a_e4hrwt_s	163	3.012	3	1.100	1	5	0.004	2.512	0.336
cancelop1mn_s	184	3.212	3	1.332	1	5	0.136	1.890	0.000
wt_heart_s	32	4.875	5	0.707	1	5	-5.388	30.032	0.000
breastwt2wk_s	158	3.734	5	1.537	1	5	-0.562	1.632	0.000
delay_dis_s	175	3.034	3	1.098	1	5	0.011	2.498	0.272
inp_survey_coord_s	176	2.989	3	1.095	1	5	-0.004	2.514	0.306
inp_survey_env_s	176	2.994	3	1.093	1	5	-0.015	2.534	0.352
inp_survey_inf_s	176	2.989	3	1.095	1	5	-0.004	2.514	0.306
inp_survey_phys_s	176	2.989	3	1.095	1	5	-0.004	2.514	0.306
inp_survey_acc_s	176	2.989	3	1.095	1	5	-0.004	2.514	0.306
inp_survey_resp_s	176	2.989	3	1.095	1	5	-0.004	2.514	0.306
dqi_s	185	3.011	3	1.098	1	5	0.003	2.500	0.253
staff_survey_s	173	3.017	3	1.102	1	5	-0.008	2.491	0.264
jundoc_s	182	3.011	3	1.102	1	5	0.003	2.497	0.253
sick_rate_s	172	3.017	3	1.105	1	5	-0.008	2.477	0.237
info_gov_s	186	3.118	3	1.064	1	5	-0.102	2.652	0.526
chi_review	90	2.422	2	0.848	1	4	0.857	2.699	0.012

Table 39: Frequency and percent distributions for transformed variables

	0	1	2	3	4	5	Total
pi_stars	8	37	86	50			181
	4.42	20.44	47.51	27.62			100
inpwt18mn_s		7	10	169			186
		3.76	5.38	90.86			100
wt15mn_s		1		185			186
		0.54		99.46			100
outwt26wk_s		2	3	181			186
		1.08	1.61	97.31			100
a_e12hrwt_s		5	33	147			185
		2.7	17.84	79.46			100
cancelops_s		4	103	79			186
		2.15	55.38	42.47			100
cancerwt2wk_s		7	67	99			173
		4.05	38.73	57.23			100
iwl_s				186			186
				100			100
cleanliness_s				186			186
				100			100
finman_s		20	15	151			186
		10.75	8.06	81.18			100
cnst_s		8			147	31	186
		4.3			79.03	16.67	100
d_esurgstd_s		41		102		23	166
		24.7		61.45		13.86	100
d_heart_s		4		22		3	29
		13.79		75.86		10.34	100
readmisn_s		63		30		73	166
		37.95		18.07		43.98	100
readm_child_s		40		64		56	160
		25		40		35	100
readm_hip_s		14		148		3	165
		8.48		89.7		1.82	100
readm_stroke_s		12		148		4	164
		7.32		90.24		2.44	100
dis_hip_s		27		115		21	163
		16.56		70.55		12.88	100
dis_stroke_s		19		138		7	164
		11.59		84.15		4.27	100
wait6mn_s		18	37	74	37	19	185
		9.73	20	40	20	10.27	100
wttarget_s		17	37	75	37	20	186
		9.14	19.89	40.32	19.89	10.75	100
outwt13wk_s		18	36	72	36	18	180
		10	20	40	20	10	100
a_e4hrwt_s		16	32	66	32	17	163
		9.82	19.63	40.49	19.63	10.43	100
cancelop1mn_s		18	37	73		56	184
		9.78	20.11	39.67		30.43	100
wt_heart_s		1				31	32
		3.13				96.88	100

breastwt2wk_s		16 10.13	32 20.25	20 12.66		90 56.96	158 100
delay_dis_s		16 9.14	35 20	70 40	35 20	19 10.86	175 100
inp_survey_coord_s		18 10.23	35 19.89	71 40.34	35 19.89	17 9.66	176 100
inp_survey_env_s		18 10.23	34 19.32	72 40.91	35 19.89	17 9.66	176 100
inp_survey_inf_s		18 10.23	35 19.89	71 40.34	35 19.89	17 9.66	176 100
inp_survey_phys_s		18 10.23	35 19.89	71 40.34	35 19.89	17 9.66	176 100
inp_survey_acc_s		18 10.23	35 19.89	71 40.34	35 19.89	17 9.66	176 100
inp_survey_resp_s		18 10.23	35 19.89	71 40.34	35 19.89	17 9.66	176 100
dqi_s		18 9.73	37 20	74 40	37 20	19 10.27	185 100
staff_survey_s		17 9.83	34 19.65	69 39.88	35 20.23	18 10.4	173 100
jundoc_s		18 9.89	36 19.78	73 40.11	36 19.78	19 10.44	182 100
sick_rate_s		17 9.88	34 19.77	68 39.53	35 20.35	18 10.47	172 100
info_gov_s		15 8.06	30 16.13	79 42.47	42 22.58	20 10.75	186 100
chi_review		5 5.56	58 64.44	11 12.22	16 17.78		90 100