2 Review of the literature

2.1 General introduction: in-hospital mortality

Although users of hospital care might consider variations in mortality rates to be of significance in their own right, the increased interest in them in recent years relates primarily to their role as indicators of broader issues in relation to the safety and quality of care provided within hospitals.

2.2 Introduction to the literature review

The narrative review that follows contains a broad introduction, and a description of the search process. Then there is an analysis of what is known about the extent of variations in hospital mortality rates, and of sources of variation; this incorporates a discussion of risk adjustment. There is a section devoted to the analysis of the relationship between variations in hospital mortality and other measures of safety and quality. Hospital mortality as an indicator is then assessed against a series of general and technical issues in relation to criteria for indicator development (Scobie et al. 2006).

2.2.1 Developments from 1860 to present

The issues around hospital mortality rates were clearly articulated in the middle of the nineteenth century (Spiegelhalter 1999). Between 1861 and 1865, the *Journal of the Statistical Society of London* published a series of articles describing hospital mortality rates, probably at the urging of Florence Nightingale. Nightingale advocated the publishing of uniform hospital statistics because these would 'enable us to ascertain the relative mortality of different hospitals, as well as of different diseases and injuries...'(Nightingale 1863).

Nightingale was very interested in the issue of quality within hospitals. She hoped that such statistics would ensure that 'As regards their sanitary condition, hospitals might be compared with hospitals and wards with wards' (Nightingale 1860). The kinds of dilemmas that the publication of such statistics would raise were also clearly understood by Nightingale, including the importance of risk adjustment for age, sex and complications (Nightingale 1863). These issues were well canvassed in the comments of Guy (1867) in relation to variations in the mortality of London hospitals.

Guy stated that 'it would be no less invidious than unjust to attribute the differing death–rates of our hospitals, in an appreciable degree, to any difference in the professional skill and ability of their professional staff, chosen, as it is, from among those members of the profession [including himself] who have already given proofs of sound training, ability and skill in practice' (Guy 1867).

Although Guy (1867) attributed the variations in hospital mortality to casemix ('…the mortality of hospitals is mainly due to causes which determine the nature and severity of the cases admitted within their walls…'), the mortality rates he quoted were not in fact adjusted for such variations, so the basis of his assertion is unclear.

Interest in variations in hospital mortality remained sporadic until the end of the 1980s, despite the unexpected findings of substantial inter-hospital variation in post surgical mortality in the National Halothane study in the USA conducted in the 1960s (Moses & Mosteller 1968). A review published in 1989 (Fink et al. 1989) could only find three articles (from 22 identified after a search) that contained any kind of adjustment for severity of illness, as well as demographic and health status issues.

At that point, the forced release by the USA Health Care Financing Administration (HCFA) of mortality rates of Medicare patients for all Medicare provider hospitals led to a major surge in interest in the analysis of hospital mortality. It appears that there was concern that the introduction by HCFA of a fixed-fee prospective payments system for Medicare patients—based on diagnostic related groups—might lead to a decrease in the quality of care provided (e.g. Stern & Epstein 1985; Iglehart 1986). Calculations of hospital mortality were a monitoring activity related to the introduction of the prospective payments system.

The public release of the HCFA information sparked considerable professional and community interest. Although subsequent studies confirmed that there were indeed variations in hospital mortality rates (e.g. Dubois et al. 1987; Chassin et al. 1989 Bradbury et al. 1991; Manheim et al. 1992; Thomas et al. 1993), a debate ensued as to the extent to which hospital level variations in mortality measures were sufficiently reflective of variations in the quality of hospital care to be broadcast to a non-professional audience, or to influence funding or purchasing decisions by insurance groups or other funders (Green et al. 1991; Hofer & Hayward 1996).

The intensity of the questioning was such that in 1993 HCFA ceased producing mortality measures. But interest in mortality measures did not decline, and as concerns have been examined and health-care providers have become more used to the release of mortality data, the frequency with which comparative risk-adjusted mortality measures have been made available to institutions and the public at large has increased year by year, and country by country.

2.3 Search method

We searched PubMed (last search updated June 2008) with a focus on studies where mortality was the primary outcome.

We searched with a variety of strategies using the following search terms: hospital mortality, review quality + risk-adjusted mortality, review risk-adjusted mortality, risk-adjusted mortality methods, risk-adjusted mortality rates, risk-adjustment methods, hospital mortality classification, history mortality measurement, quality risk-adjusted mortality rates, quality + risk-adjusted mortality rates, hospital standardised mortality ratios.

We focused on studies that compared whole of hospital mortality rates and related the results to any evidence of quality and or safety. Although we did find many studies looking at only a single condition—such as acute myocardial infarction (AMI), coronary artery bypass grafting (CABG), and pneumonia—or only one hospital, those were not our primary interest. We aimed our review at studies that compared at least two hospitals. Studies that looked at mortality through the lens of organisational/structural variables, nurse–patient or physician–patient ratios, and public versus private funding were not our prime focus.

In our search we paid particular attention to national mortality rate reporting that has recently been undertaken in the United Kingdom, United States, Canada and Holland.

For all studies, the authors decided final inclusion/exclusion by discussion and consensus.

2.4 Considerations in the development of mortality as an indicator

2.4.1 Random and systematic variation

Before any attempt is made at interpreting or using hospital mortality data, a basic issue needs to be understood and responded to.

Hospital mortality is a special case of a more general issue related to the analysis of variations in the outcomes of any intervention (Thomas & Hofer 1998) when the factors involved are both systematic and random. As with any other outcome in biomedicine, variations in mortality outcomes following hospitalisation can be thought of as having at least two components:

- systematic variations in factors that may influence [mortality] outcomes; those variations being assumed to relate to the quality and effectiveness of the interventions that affect the outcome in question
- random variations.

The random variations may have a variety of origins. There are random variations in the moment-by-moment effectiveness of biomedical interventions, even when they are optimally applied. There are random variations in the interaction between optimally applied interventions and the immediate states of the people to whom those interventions are applied, and random variations during attempts to implement evidence-based interventions (the systematic consequences of the longer term characteristics, or traits, of those people are best thought of as confounders of systematic variation and are considered below).

In biomedical research, the uncertainties due to random variations are optimally dealt with by a process of randomisation. When patients are randomly allocated to the settings or interventions of interest, the presence of a systematic effect is confirmed by assessing the magnitude of differences in outcomes between sites or interventions, taking overall variability into account. The fundamental analytical question is whether the observed differences are so large that they are unlikely to have occurred by chance alone.

Hospital mortality measures are measures of outcomes in the usual care provided by hospitals. There is no possibility of random allocation of patients to different sites. The question of whether observed differences are so large that they are unlikely to have occurred by chance can only be assessed by comparing the outcomes for a patient or group of patients treated in any one hospital against a hypothetical outcome that might have occurred if the patient(s) had undergone treatment elsewhere.

The most straightforward way to do that would be to assess the average outcome across the population being assessed and use that to calculate the expected outcome (and confidence limits around the value) for the number of patients treated at any one hospital. The observed (actual) and expected values for the numbers of patients treated would be compared and a decision made as to whether any hospitals stand out as being ‘extreme’ in terms of differences between observed and expected outcomes. However, a simple comparison on that basis is likely to be both inaccurate and misleading.

Patients are non-randomly allocated (and self referred) across institutions. The use of crude averages ignores patient-level differences between institutions that might systematically influence outcomes. These confounding factors, combined, may be described as variations due to the clinical, demographic and casemix differences between patients present at the point of arrival in hospital (V_C). In which case, total variation in in-hospital mortality (V) comprises:

- systematic variations in factors influencing mortality outcomes; those variations being assumed to relate to the quality and effectiveness of the interventions that affect the outcome in question (V_Q)
- variations due to the clinical, demographic and casemix differences between patients present at the point of arrival in hospital (V_C)
- random variations (V_R).

In most studies of hospital mortality, efforts are made to discount V_C before assessing the magnitude of any inter-hospital differences (Thomas & Hofer 1998). The measurement of V_C for this purpose is usually described as risk adjustment because pre-existing patient-level factors influence or confound any other institutional-level factors that might influence the risk of dying in hospital. There is also the possibility that there are some confounding factors related to the characteristics of the functional catchment areas of hospitals that are not captured in existing individual-level measures, and that need to be accounted for by inserting measures of social disadvantage into analyses (Jarman et al 1999). Whilst there is disagreement as to whether such influences should or should not be adjusted for, the question of the influence of catchment population measures on in-hospital mortality in the Australian context is assessed empirically in this project (Section 5.9).

Much of the criticism of the release of the HCFA mortality studies of the 1980s (Rosen & Green 1987; Berwick & Wald 1990; Green et al. 1991) related to the fact that the risk adjustment was confined to the impact of each patient's principal diagnosis and four secondary diagnoses, and demographic factors of age, sex, race, and whether the patient had been transferred from another hospital. Critics argued that this was too simplistic to adequately adjust for patient-level variations between institutions (Green et al. 1991).

2.4.2 Mortality at what point: in-hospital, 30 days after discharge, or longer?

Another common complaint in the literature following the release of the HCFA data was that many of the effects of hospital care do not become evident until after patients leave hospital. Also, if studies of variations in mortality rates were to be confined to deaths during hospital stays, hospitals might be tempted to discharge poor prognosis patients to minimise in-hospital mortality (Omoigui et al. 1996)

By linking hospital data with relevant information from death registers, a number of investigators have assessed the relationship between mortality during hospital stay and mortality 30 days after discharge (Jencks et al. 1988; Chassin et al. 1989, Rosenthal et al. 2000) or longer (Fleming et al. 1991; Garnick et al. 1995). Inclusion of deaths in the thirty-day period after discharge appears to be sufficient. After an exhaustive study, Garnick et al.

(1995:693) concluded that 'mortality occurring after 30 days has little to do with hospital-specific effects…'

As may be expected, mortality up to 30 days after discharge is tends to be similar to in-hospital mortality (e.g. Rosenthal et al. 2000), but this is not necessarily so, and variations have the potential to be informative. Assessing mortality up to 30 days after leaving hospital provides the opportunity to assess effects of variations in discharge policy (Jencks et al. 1988) and of immediate post-discharge care.

Whilst it may thus be preferable to assess mortality in a manner that includes deaths up to 30 days after discharge, it is not always feasible to do so, and the gain in precision by taking account of mortality after discharge has to be traded against the greater complexity involved in linking hospital administrative information with other registry data (Krakauer et al. 1992). However, developments in population-level data linkage capabilities, such as the Western Australian Data Linkage System and the work of the Centre for Health Record Linkage in NSW, are reducing this barrier and will offer further opportunities in the future.

2.5 Model development

2.5.1 What variables to include in risk adjustment

Demography and diagnosis

The risk-adjustment hypothesis is that observed rates of in-hospital mortality will be systematically influenced by the characteristics of patients on arrival at the hospital.

It seems reasonable to assume that the risk of death during a hospital stay is likely to be influenced by factors such as age, sex, primary clinical diagnosis and secondary or complicating diagnoses present at admission. Information on these types of factors is commonly collected within administrative data sets—that is, within information about individual patients collected by hospitals for internal and external administrative reasons and mandatory reporting requirements. Hospital-level administrative data sets in Australia and elsewhere also commonly contain information about arrival and discharge dates, home address, source of referral, whether the admission was as an emergency or planned, and the nature of discharge. Information about ethnicity may or may not be available, along with other jurisdiction-specific information.

Severity

Administrative data sets do not usually contain much information about the severity of the principal diagnosis, though this varies between diagnoses. For example, Australian data coded according to the International Classification of Diseases Australian Modification of the ICD (ICD-10-AM) do not usually provide information on the severity of an uncomplicated case of community-acquired pneumonia over and above the diagnosis itself. The same classification does, however, distinguish depressive episodes as mild, moderate, severe and severe with psychotic symptoms, and liver lacerations as minor, moderate and major.

Severity is neither a simple nor a uniform characteristic, nor easily or uniformly assessed. For instance, the severity of heart disease may be inferred from physiological or medical imaging data reports, whereas the severity of schizophrenia is best determined by clinical judgment.

Institutional characteristics

Many administrative data sets that report patient-level data also characterise the reporting institutions in some way. The basic requirement is for a field in patient-level records that records the treating hospital[1]. This is particularly important when a data set contains outcomes from both large principal referral hospitals, and small institutions. The case loads of small hospitals are often quite different from those of tertiary institutions. Patients in smaller institutions can appear to be at lower risk than patients in larger institutions, even after risk adjustment. However, it is not appropriate to assume that the smaller hospital could achieve similar types of outcomes if they were confronted with the kinds of patients that tertiary institutions have to deal with. A low-risk hospital is only low risk for the kinds of cases it is familiar with (Shahian & Normand 2008). So, institution type is a relevant issue when making comparisons. Risk adjustment itself is, however, best undertaken at the level of the patient, not the institution (e.g. Hadorn et al. 1993).

2.5.2 Logistic regression and risk adjustment

The 'mechanics' of risk adjustment—once potential risk modifying factors have been identified—are well established. Taking hospital mortality as the dependent variable, the influence on outcome of various independent variables (or contributors of mortality) is assessed by means of logistic regression: the appropriate analytic strategy for binary (survive/dead) outcomes. Logistic regression allows development of a linear equation for the log (odds) of a positive outcome. The log (odds) increases by the magnitude of the coefficient for each unit increase in the independent variable. For example the log (odds) of a positive outcome for male versus female increases by the coefficient for sex, if male is coded 1 and female is coded 0.

The exponentiated coefficients can then be interpreted as the change in the odds of a positive outcome for a unit increase in the associated independent variable (i.e. covariate).

The coefficients from logistic regression can also be applied to create a predicted probability of an outcome of interest (i.e. death) for each individual in the data set. The probabilities for each particular pattern of covariate values effectively create a set of reference weights that relate to the population of hospitals as a whole, enabling standardisation of each individual hospital to a reference hospital population. The aim is to profile how the results for a particular hospital compare with what would be expected if that hospital functioned in a way that was typical for the whole population of hospitals studied.

2.5.3 Logistic regression, indirect standardisation and HSMR

Each patient in any one hospital will survive or die. The sum of all the deaths divided by the total number of hospital separations is the crude in-hospital mortality rate for that hospital. By calculating the probability that any one patient in a population of patients will die (or survive) using the logistic regression coefficients and covariate values relevant to that patient, it becomes possible to compute the standardised mortality rate for that institution; that is, a mortality rate that is adjusted for its casemix.

[1] Some private hospitals are not identified as separate establishments in the Australian hospitals data available for this project (see Appendix 5 Data issues).

Indirect standardisation of hospital mortality rates is the term given to the comparison of the observed mortality rates against the expected rates as generated from the study of all the patients within the hospital populations studied. Those expected rates become the denominator of the ratio of observed to expected outcomes (O/E). A ratio value less than 1 is favourable and a ratio of greater than 1 unfavourable. When the ratio is multiplied by 100 the convention is to describe that value as the Hospital Standardised Mortality Ratio (HSMR) (Jarman et al. 1999).

Although the computational method used in risk adjustment for the calculation of hospital level HSMRs is now fairly settled, the range of contributing variables that might be included in the regression equation is almost without limit. In practice, there is an emerging consensus on which variables to include in studies analysing the majority of deaths occurring within hospitals (as distinct from studies dealing only with deaths of specific types).

2.5.4 Narrowly focused or broad studies

Which patients should be included in the study of mortality rates? Should the study be narrow focused or more broadly based?

Over the years, studies have examined mortality rates in single conditions, small groups of diagnoses with high predicted short-term mortality, patients from diagnostic groups in which the majority of deaths within hospital occur, or all patients treated with a limited number of exclusions. Despite the substantial potential differences involved, there has been little discussion of the rationale behind any one choice, though there are some practical issues to be considered.

Studies of hospital mortality easily accumulate very large numbers of subjects; for example, the national data set for all separations from Australian hospitals in the financial year 2005–06 contains some 6 million individual records. Data sets from countries with higher populations, such as the UK or USA, will be proportionately larger.

The surge in interest in hospital mortality began in the late 1980s. Although it is not explicitly discussed in the literature, very large data sets were not as easy to handle at that time as they are now. The greater expense then of acquiring access to sufficient computing power would have been a consideration in favour of opting to confine analyses to subsets of the whole population of patients treated in hospitals.

A limited number of clinical conditions accounts for the majority of deaths within hospital. When analyses examine mortality rates within the diagnosis groups that account for 80% of all hospital deaths, clinical diagnoses—albeit somewhat simplified or consolidated—can be included directly within risk-adjustment methods (e.g. Kahn et al. 1990). Once studies encompass all deaths within a population of hospital patients, then some means of aggregating diagnoses into larger groups is required because the numbers of individual diagnoses are just too large for all those diagnoses to be individually included in risk-adjustment computations.

In all studies, provision is made to exclude those patients for whom death in hospital is integral to the service provided. Strategies have been developed to deal with palliative-care-type hospital separations (CIHI 2007). In Australia, palliative care is designated within administrative data sets as a care type that can only be provided in a designated Palliative Care service. It is straightforward to exclude such patients. In settings where that is not possible, other arrangements are required to deal with potential palliative-care issues, such as excluding patients with a primary diagnosis of cancer (e.g. Lakhani et al. 2005).

Restricting the analysis of mortality to a small number of conditions may be relevant if there is a strong interest in linking mortality outcomes with specific process measures. Otherwise, a broader sample of in-hospital deaths is likely to provide a more representative population for analysis. The case for confining a more broad-based analysis to the higher risk diagnoses that account for 80% of deaths—instead of all in-hospital deaths—has not been formally argued, and relates more to convenience and the capacity to include primary diagnoses as they stand within the risk-adjustment process, than to other issues of substance. The analyses further include high-risk diagnoses, low-risk diagnoses, and all causes of mortality.

2.5.5 Summary measures of model performance

The underlying rationale for logistic regression is that the risk of an event in relation to risk factors falls along a logistic curve. The s-shaped logistic curve is shown below, where 0 on the y-axis is alive, and 1 the outcome dead, and the values between are the probabilities of the outcome.

Logistic regression analyses are mathematical models that attempt to fit the data to the logistic curve. Commonly asked question of such models are 'How good is it? What is its predictive validity—how well does the model account for the actual variation in patient-level risks (Shwartz & Ash 2003)?'

There is some controversy in the technical literature about what, if any, are the best summary measure to use to answer such queries. There are two issues to be considered: null model and goodness of fit.

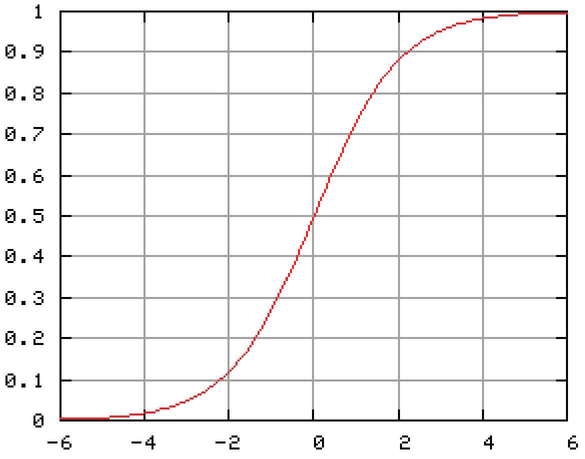


Figure 1: Example of an s-shaped logistic curve

2.5.6 Null model

Firstly, do the models created improve upon the 'null model'? Say we are interested in examining the mortality at St Elsewhere—one of a population of hospitals for which in-hospital mortality is being studied. If there is no opportunity to risk adjust by reference to additional variables, the only way to define the expected numbers of deaths in St Elsewhere is to take the average death rate for all hospitals and apply that rate to the total number of patients treated in St Elsewhere, deriving a predictive 'null' model using that information alone.

If patient-level confounders are important, adding them to the model will improve predictive power over a model with no other adjustment variables. Whether any improvement is statistically significant may be tested by means of a likelihood ratio (LR) test. LR tests examine the predicted probabilities of living among those who lived, and the predicted probabilities of dying amongst those that died. Better models have higher LRs (i.e. more of the living were predicted to have lived, and more of the dying were predicted to have died).

2.5.7 Goodness of fit

Goodness of fit is a somewhat different question. The issue is not 'does it fit better than the null model?' What is being asked is 'how well does the model fit?' It may be better than chance, but how strong is the relationship?

The challenges posed by such questions are best appreciated by comparing logistic regression models with the more straightforward measures generated for linear relationships. There, the relationships between the dependent and independent variables can be considered as potentially falling along a straight line. When increases in the independent variables are perfectly mirrored in increases in the dependent variables, an equation linking the two groups of variables will predict 100% of the variability in the values of the dependent measure. If there is no link at all, then the equation will predict 0% of the variability. By calculating the R^2 statistic, the percentage of variability explained by the equation can be calculated (i.e. how closely do the points in the scattergram linking independent and dependent relationships fit to a straight line?).

R^2 (or pseudo R^2, a related measure) can be calculated in logistic regression, but the results cannot be interpreted in the same way as in a linear regression. The issue of interpretation goes back to the fact that a logistic regression is an attempt to predict the degree to which a group of variables (such as age, sex, and admission status) predict a binary (alive/dead) outcome, not a graded one. Conceptually, the analytical question asked is 'does a risk-adjusted equation produce a result that, when applied to a population, sharply separate the population who are alive at discharge from those that die in hospital, with limited overlap between the two groups?' The problems with interpreting R^2 as a measure of 'model fit' for logistic regressions were summed up (Schwartz & Ash 2003) in a discussion of the publication of CABG data in New York (Chassin et al. 1996).

> 'In logistic-regression models in which the overall mortality rate ranges from 2 to 4 per cent, however, R2 is almost always less than 0.2. This limitation arises from the nature of logistic regression, in which the dependent variable must have one of only two values (in this case survival or death). When the differences between actual and predicted mortality rates is calculated for each person (as part of the calculation of R2) no matter how accurate the prediction is, the difference between the predicted value and the

> observed value for the mortality will be large, because the observed mortality must be either 0 or 1, and the prediction is a proportion between 0 and 1.' (Chassin et al. 1996: 396–7).

Using changes in R^2 to assess the impact of adding or subtracting variables within a logistic regression model remains valid, however, because this is using it in a variable-by-variable comparison, rather than in an attempt to provide a single statistic against which to assess model fit.

2.5.8 The c-statistic

A better measure of discrimination is the c-statistic, which also equals the area under a receiver-operator curve (ROC). The c-statistic has a number of definitions, but one is as follows.

> 'Within a population, take all the possible pairs in which one patient dies and the other survives. Assign a probability of death for each patient in each pair. The c-statistic equals the proportion of cases in which the predicted probability of death is higher for the patient who died than the patient who lived. When the probability is tied, the assigned value is one half—that is, there is a 50:50 chance of being right or wrong. So when models have no ability to discriminate—that is, to truly assign a probability of death while minimising false positives—the c-statistic is 0.5. Although there are no absolute hard and fast rules, models generating a c-statistic value below 0.7 are considered to be poorly discriminatory, models with a c-statistic 0.7–0.8 are more adequate, and above 0.8 a good discrimination' (Aylin et al. 2007).

As will be shown below, many risk-adjustment models for mortality have c-statistics in the range 0.8 and above.

2.5.9 Risk adjustment across the range of predicted probabilities

Many studies of hospital mortality will involve patients across a wide range of risk. One method for assessing the robustness of risk adjusters across the whole range is the Hosmer–Lemeshow method (Hosmer & Lemeshow 2000).

Patients are divided up into deciles of predicted risk and the observed and expected values of mortality (derived from applying the coefficients of the logistic regression to the populations) calculated for each decile. The distribution of the deviations within each decile follows the chi^2 distribution, and the model is accepted if the observed deviations or differences are *less* than would be expected by chance. Despite the elegance of this method, the Hosmer–Lemeshow test, like all chi-square tests, is sensitive to sample size, and may not be suitable for studies with large samples (Schwartz & Ash 2003; Aylin et al. 2007). The direct comparisons between observed and expected values at deciles of risk may be of considerable interest (Aylin et al. 2007), and may provide insights into the impact of risk adjustment without further analysis.

2.5.10 Calibration

An entirely different issue is that of calibration. Because the risk-adjustment process begins with the calculation of an expected or average outcome, the overall observed and expected outcomes will be identical, because the expected is the average of the observed.

When a risk-adjustment equation is calculated in one population and then applied to a quite different one, the calculated expected number of deaths will not necessarily be the same as the observed. The question arises as to whether the expected results should be calibrated, or adjusted in some way, so that the overall expected and observed values resemble each other. A number of calibration methods have been suggested in the literature (see DeLong et al. 1997) but, although this is a theoretically important issue, and would need to be considered carefully if there were any attempt at a
cross-national comparison of HSMRs, it has only received limited empirical study to date.

So, in summary, there are a variety of measures that can be used to assess the robustness of a risk-adjustment process for binary outcomes, but none give a simple answer to the question 'how good is the fit?'

2.6 Inter-hospital variation and risk-adjustment models

2.6.1 Hospitals differ

After interest in variations in hospital mortality picked up following the publication of the HCFA data, the fact of highly statistically significant variations in in-hospital mortality rates have been confirmed in every country where they have been studied (e.g. Chassin et al. 1989; Kahn et al. 1990; Jarman et al. 1999; CIHI 2007; Heijink et al. 2008), in public and private hospitals alike (Devereux et al. 2002).

2.6.2 Risk adjustment—administrative data sets

Table 1 provides a listing for the R^2 and c-statistic values for a variety of reports of risk-adjustment models, and the values for the areas under the ROC where provided.

Numerous reviews of the outcomes of risk adjustment using administrative and other data sets have been published over the years (e.g. Hadorn et al. 1993; Iezzoni 1997a; Thomas & Hofer 1999; Powell et al. 2003; Daley et al. 2003), and it is now possible to draw some overall conclusions.

Administrative data sets contain a restricted amount of information at the patient level. Demographic information, mode of admission (emergency or elective, transfer from other care facility or direct) and duration of admission, care type, mode of discharge, principal and secondary diagnoses, surgical procedures, and institutional identifiers are almost always available. The Australian administrative data sets separate types of care into acute, rehabilitation and palliative care. Information about previous admissions and linkage across hospital and community services are less common.
(The spreading availability of data linkage facilities in Australia is overcoming this limitation.)

In the UK, there have been particular difficulties relating to the use of multiple consultant-completed episodes within a single admission that have had to be overcome (Jarman et al. 1999), but that is not a widespread problem outside the UK.

The most important changes over the years have related to the increase in the number of primary and secondary diagnoses that are contained within administrative data sets, with restricted numbers (e.g. in the HCFA) now commonly replaced by more exhaustive enumerations in many countries. For example, the current Australian National Morbidity Collection allows for the reporting of one primary and
49 secondary diagnoses, and up to 50 procedure codes.

A more subtle issue relates to the notion of what constitutes the principal diagnosis for a patient. In most systems that derive from the Medicare-derived USA prospective payment systems, the convention is that the primary diagnosis is the diagnosis that, after study, was the primary condition leading to hospital admission. But in the large USA Department of Veteran Affairs system, it is the condition primarily responsible for the length of the hospitalisation (Daley et al. 1997, Iezzoni 2003b). In Australia, principal diagnosis is defined within the National Health Data Dictionary as 'The diagnosis established after study to be chiefly responsible for occasioning an episode of admitted patient care, an episode of residential care or an attendance at the health-care establishment' (AIHW 2006). The specification of the principal diagnosis may have an important bearing on the risk rating of each patient.

Although concerns have frequently been raised over the accuracy of coding of diagnoses (e.g. Iezzoni 1997a, Scott & Ward 2006) those concerns have tended to become less prominent in recent years, as countries have become familiar with the work of professional coders, and as work on coding standards and coding practice has become increasingly refined.

Within the National Hospital Minimum Dataset, only a small percentage of cases are recoded due to an error (AIHW 2007), with most errors being in the direction of 'up-coding' in the direction of increased complexity, which would tend to reduce any measure of hospital mortality because observed mortality would tend to be less than expected mortality in those cases.

Studies of the outcome of risk adjustment via administrative data systems have been reviewed on a number of occasions (e.g. Hadorn et al. 1993; Iezzoni 1997b; Thomas & Hofer 1999; Powell et al. 2003; Daley et al. 2003), and a number of different methods for combining the information within administrative data sets have emerged, including a number of proprietary methods developed in the USA (e.g. the APR-DRG system, Disease Staging). However the R^2 model statistics reported in Table 1 have not varied from the 0.2 to 0.3 levels reported by Hadorn et al. in 1993, and the c-statistic levels continue to typically range from 0.7 to 0.8 or slightly above.

In the next section, the addition of clinical elements to risk adjustment is discussed. Because it stretches across both administrative and clinical risk-adjustment methods, a discussion of the integration of comorbidities in risk adjustment is undertaken further on.

2.6.3 Risk adjustment—the addition of clinical factors

Clinicians make judgments based on the clinical characteristics of their patients, so it would seem axiomatic to those practitioners that outcome predictions that include clinical information would be an improvement over those that do not. It is not surprising, therefore, that considerable effort has gone into the search for clinical elements to test in risk-adjustment models. The rationale for those attempts was put elegantly by Hadorn et al. (1993: 1–2), 'Statistical prediction models rely on the same clinical and demographic factors (e.g. age, blood pressure) used by clinicians to arrive at prognostic judgments. Unlike clinicians, however, models assign explicit weights to these factors based on their observed statistical association with the outcome of interest (e.g. mortality) in some sample of patients. As a result, prediction models render precise (if not always accurate) predictions of outcome or diagnosis.'

The simplest of these strategies has been to model physiological data (e.g. blood pressure in the first 48 hours of stay) or laboratory test values (for potassium, haematocrit, and so on), and include them as confounders within models to risk adjust mortality data. Then there are strategies that generate condition-specific measures combining laboratory and clinical elements, using guidance from clinical panels or other sources of clinical advice to choose from among candidate variables, extracted from case notes by trained reviewers, to test in risk adjustment.

Finally, there are proprietary services (e.g. MedisGroups) whose trained personnel (commonly nurses) review case records and extract and tabulate many different features of interest that can be tested in risk-adjustment studies. Iezzoni (1997a, 1997b) describes the origin of one of the most widely used of these methods, the MedisGroups listing of key clinical findings, in the observations made by two physicians from Saint Vincent's Hospital in Worcester, Massachusetts, after participating in the morning reporting process of medical residents. These observations eventually became the initial list of what are now hundreds of key clinical findings.

Table 1 provides a selection of the model parameters from risk-adjustment models using a variety of clinical risk parameters. Although many of them do improve on the R^2 for the risk-adjustment methods based on administrative data, the gain is often modest.

Given the variety of administrative and clinical risk-adjustment methods that have emerged, the series of studies of Iezzoni and colleagues conducted during the mid 1990s are particularly important (the outcomes are tabulated in Table 1, and overall outcomes summarised in Iezzoni (1997a, 1997b). These researchers compared a wide variety of risk-adjustment methods using a single data set as the test or trial data source. They directly compared a wide variety of risk-adjustment methods for AMI, coronary by-pass artery grafting, pneumonia or stroke, and compared five of the methods on all four diagnostic groups.

The risk-adjustment methods studied included Disease Staging, All-Patient Refined Diagnostic Related Groups (APR-DRGs) and Patient Management Categories (PMC): all three being proprietary risk-adjustment methods that made use of discharge abstracts (i.e. administrative data sets). MedisGroups and the APACHE 111 system represented risk-adjustment methods that made use of physiological and or clinical data.

The results were clear. Although risk adjustment is necessary for valid comparison of hospitals or groups of hospitals, no particular method stood out as preferable. Whilst the methods tend to agree on high and low mortality outliers, no one method provided markedly more specific and consistent discrimination than the others.

Table 1: Risk-adjustment-model outcomes

Year	First author	Condition(s)/severity adjustment	R^2	C	ROC
1985	Knaus	ICU—Apache 1	0.31		0.851
		ICU—Apache II	0.319		0.863
1990	Keeler	stroke—Apache II	0.30		
		pneumonia—Apache II	0.26		
		myocardial infarction—Apache II	0.22		
		heart failure—Apache II	0.12		
1991	Knaus	ICU—Apache III on initial day	0.41		0.90
1992	Krakauer	multiple—demographic model		0.64	
		multiple—claims model		0.84	
		multiple—clinical model		0.90	
1994	Hannan	CABG		0.79	
1995	Green	CABG	0.073		
1995	Romano	AMI—model A		0.766	
		AMI—model B		0.844	
		Lumbar diskectomy—model A		0.722	
		Lumbar diskectomy—model B		0.73	
		Cervical diskectomy—model A		0.702	
		Cervical diskectomy—model B		0.744	
1996a	Iezzoni	pneumonia—medisgroups or	0.13	0.81	
		pneumonia—medisgroups exp	0.19	0.85	
		pneumonia—physiology 1	0.10	0.78	
		pneumonia—physiology 2	0.15	0.82	
		pneumonia—body systems count	0.05	0.71	
		pneumonia—comorbidities index	0.06	0.74	
		pneumonia—disease staging	0.13	0.80	
		pneumonia—PMC severity score	0.11	0.79	
		pneumonia—AIM	0.05	0.73	
		pneumonia—APR DRGs	0.10	0.78	
		pneumonia—PMC RIS	0.1	0.78	
		pneumonia—R DRGs	0.28	0.83	
		pneumonia—age sex interact only	0.03	0.67	
		pneumonia—age sex interact, DRG	0.04	0.71	
1996b	Iezzoni	AMI—medisgroups or	0.17	0.80	
		AMI—medisgroups exp	0.23	0.83	
		AMI—physiology 1	0.18	0.82	
		AMI—physiology 2	0.23	0.83	
		AMI—disease staging	0.27	0.86	
		AMI—PMC severity score	0.18	0.82	
		AMI—comorbidity index	0.06	0.70	
		AMI—APR DRGs	0.20	0.84	
		AMI—R DRGs	0.15	0.80	
		AMI—age sex interacted	0.05	0.69	

(continued)

Table 1 (continued): Risk-adjustment-model outcomes

Year	Author	Condition(s)/severity adjustment	R^2	C	ROC
1997a	Iezzoni	AMI—medisgroups	0.227	0.83	
		AMI—Physiology score	0.229	0.83	
		AMI—disease staging	0.27	0.86	
1997b	Iezzoni	AMI—PMC severity score	0.176	0.82	
		AMI—APR DRGs	0.198	0.84	
		CABG—Medisgroups	0.036	0.73	
		CABG—Physiology score	0.028	0.72	
		CABG—Disease staging	0.069	0.77	
		CABG—PMC severity score	0.079	0.8	
		CABG—APR DRGs	0.066	0.83	
		Pneumonia—Medisgroups	0.19	0.85	
		Pneumonia—Physiology score	0.149	0.81	
		Pneumonia—disease staging	0.132	0.8	
		Pneumonia—PMC severity score	0.115	0.79	
		Pneumonia—APR DRGs	0.101	0.78	
		Stroke—Medisgroups	0.265	0.87	
		Stroke—Physiology score	0.242	0.84	
		Stroke—Disease staging	0.112	0.74	
		Stroke—PMC severity score	0.101	0.73	
		Stroke—APR DRGs	0.105	0.77	
1997	Silber	Adult surgical—Medisgroups full model		0.92	
		Adult surgical—without severity score		0.83	
		Adult surgical —without everity/emergency		0.74	
1997	Pine	AMI, cerebro, CHF, pneumonia—admin			0.75–0.87
		AMI, cerebro, CHF, pneumonia—clinical			0.86–0.87
1997	Khuri	Non-cardiac surgery—10 variables		0.87	
		Non-cardiac surgery—44 variables		0.89	
1998	Polanczyk	CHF		0.83	
1999	Ansari	Prostatectomy	0.24	0.89	
2001	Austin	AMI			0.775
2003	Tekkis	Gastrooesphageal cancer			0.78
2003	Reed	CAB—Parsonnet/recalibrate		0.752–0.805	
		CAB—Canadian/recalibrate		0.693–0.755	
		CAB—Cleveland/recalibrate		0.748–0.769	
		CAB—New York/recalibrate		0.735–0.768	
		CAB—Northern New England/recalibrate		0.772–0.803	
		CAB—New Jersey/recalibrate		0.787–0.839	
2005	Geraci	CABG		0.698	
2005	Gordon	Non-cardiac surgery		0.65–0.83	
2007	Aylin	isolated CABG, AAA, colorectal			0.66–0.803

2.6.4 Over-fitting

The Iezzoni study touched on an important issue in relation to risk adjustment based on clinical parameters. Risk adjustment involves assessing the extent to which patient-level parameters—present at the point of admission—predict an outcome at a future point. The closer a risk-adjustment model is tailored to a particular condition, or to a particular clinical setting, the less likely it is be as precise when applied to other conditions or other settings. There is no intrinsic reason why a risk-adjustment method that is tailored to predict the outcome of one condition, such as myocardial infarction, should predict the outcome of another condition, such as pneumonia, because the physiology, pathology and the range of potentially beneficial interventions are quite different.

Statistically, the risk of adjusting too closely to a particular casemix, is called over-fitting. It is assessed by means of cross-validation measures, but the problem of over-fitting represents a natural ceiling for the development of clinical risk-adjustment methods for studies of mortality across a wide range of patients. Risk-adjustment methods that have been developed on specific patient groups, or within specific clinical settings, will lose precision when applied across a broader range of patients and settings. This reinforces the utility of risk-adjustment methods that make use of the more general information in administrative data sets.

One simple test for over-fitting is to divide a data set into a test set and a confirmatory set. When the model developed with the test set is fitted to the confirmatory set, if the precision deteriorates markedly with the confirmatory set, over-fitting is likely to have occurred.

2.6.5 Further comparisons between risk adjustment from administrative and clinical databases

In an important recent study, Aylin et al. (2007) compared the discriminatory capacity of risk-adjustment models for in-hospital mortality derived from an administrative data set with models based on clinical databases compiled by professionals.

The clinical databases were compiled by the Society of Cardiothoracic Surgeons, the Vascular Surgical Society of Great Britain, and the Association of Coloproctology of Great Britain and Ireland. The conditions whose mortality was recorded were isolated CABG, repair of abdominal aortic aneurysm (AAA), and colorectal excision. The administrative data set was the UK hospital episodes statistics, with the completed consultant episodes that comprised each admission merged together.

The authors calculated the c-statistic for both a simple model derived from the administrative data (just the year of procedure, age and sex), and more complete models with primary and secondary diagnoses, method of admission, Charlson index for secondary diagnoses (see below), and socioeconomic deprivation. The models derived from the administrative data sets were compared in relation to discriminatory power against the published results of risk-adjusted models using the clinical data in the database, as generated by the holders of the databases.

The results clearly demonstrated that the models based on administrative data were as successful in discriminating cases as those derived from the clinical databases. For the repairs of AAA and colorectal excision for cancer, the models based on the administrative data showed better discrimination, and for isolated CSBG, the c-statistic was only different by 0.02.

Models derived from administrative data systems have also proved to be adequately discriminatory in a study of post surgical outcomes in the Department of Veteran Affairs surgical clinical improvement program (Geraci et al. 2005, Gordon et al. 2005).

2.6.6 The Charlson Index

Although biomedical knowledge and evidence-based practice are often derived from studies of isolated clinical disorders, patients themselves will commonly suffer from a mixture of conditions. This is increasingly important as the population ages. So risk-adjustment methods need to reflect that complexity. The dilemma is that there are so many potential individual and combined clinical comorbid confounders, that some method of data reduction or simplification becomes necessary if comorbid complexity is to be included in risk adjustment for mortality or morbidity.

In 1987, Mary Charlson and colleagues (Charlson et al. 1987) published a paper describing an index—since widely known as the Charlson Index—in which groups of clinical conditions were assigned numerical weights whose additions combined to generate an interval score that predicted increasing likelihood of death over a 1 year or longer period. The original paper described a score with values from 1 to 16.

There is now a very extensive literature relating to the use of the Charlson index as a measure for predicting mortality in many settings, and it soon became apparent that it was a useful method for grouping comorbidities in hospital mortality studies (Iezzoni et al. 1996a; Polanczyk et al. 2002; Romano & Mutter 2004; CIHI 2007; Heijink et al. 2008; Aylin et al. 2007).

Although the original version was in the ICD-9 diagnostic system, it has been converted to the ICD-10, (Sundararajan et al. 2007) with no loss of precision.

Computerised systems exist for grouping secondary diagnoses in administrative data systems, such as the Australian National Hospital Morbidity Collection, into their respective Charlson group. In addition, the widespread use of the Charlson groups for the development of risk-adjustment models for hospital mortality studies makes it clear that the groups within the Charlson index are the de-facto standard method for grouping complicating conditions both for studies of specific conditions, or broad-based measures.

Although the Charlson index groups conditions into groups of increasing ‘severity’, and aggregates those groups into an interval score that can range from 1 to 16, most studies of hospital mortality have truncated the score. In an unpublished study of hospital mortality in South Australia in 2002 (Ben-Tovim 2002), the score was truncated at 5. In the Canadian study described above (CIHI 2007), it was capped at 2, and so on. An alternative to the identification of regression coefficients related to the score assigned to the comorbidity is to aggregate the comorbidites into their Charlson group, then insert the groups into the logistic regression, and generate a group-specific coefficient (Polanczyk et al. 2002). That was also the method used in the unpublished South Australian study (Ben-Tovim 2002). When used in that way, the coefficients cannot be applied to a different population of patients without testing for over-fitting. The Charlson index in its various guises continues to be developed as a valuable tool in risk adjustment.

2.6.7 Summary of risk adjustment and hospital mortality

When studies of comparative hospital mortality are presented to clinicians, one of two stereotypical reactions often occurs. If the hospital or service involved scores 'well', then satisfaction is taken with the outcome. If the hospital or service scores 'poorly', then doubt is likely to be expressed about the data and method used, focusing on whether the method has adequately accounted for the 'difficulty' of the institution's casemix. The discussion of risk adjustment in this section has been provided with this in mind.

Some conclusions can be drawn from the sections above. Firstly, real and substantial differences can be found between hospitals in relation to in-hospital mortality. The differences are not affected greatly by whether measurement is restricted to deaths during hospital stays, though it is better to include deaths soon after discharge.

Attempts to create a level playing field for inter-institutional comparisons have their problems. There are limits to the precision of existing risk-adjustment models. Models can be developed that have acceptable discriminatory power overall, but are poor predictors of individual outcomes. This is not simply a technical problem. As practising clinicians will acknowledge, their accuracy in predicting survival or death during any one hospital stay for an individual patient who is not clearly terminally ill is limited, even in the case of the most severe illness. Survival 'against the odds' is a driving force for much clinical effort, and there are countless patients and their families who have enjoyed extra years of life as a result of those efforts. The limits of statistical methods are the limits of our understanding of the nature of illness itself.

It is also clear by now that the early concerns about the limitations of administrative data systems are unfounded. Contemporary administrative data systems—professionally extracted and coded, with a wide variety of primary and secondary diagnoses—are an acceptable source for further study of the causes of variation in hospital mortality, and there is little difference in terms of discriminatory power between models derived from them and models derived from clinical databases
(e.g. Smith 1994). This is reassuring, because the cost and complexity of extracting clinical, or even simple laboratory, information on a large scale from existing record systems on a national scale in countries such as Australia are prohibitive. This is true even in countries such as the USA, where, as Birkmeyer et al. (2006: 417) put it in 2006:

> 'Although it is not clear whether our results would have differed if we had access to detailed clinical information for better risk adjustment, this question may be moot from a practical perspective. With the exception of cardiac surgery, clinical data for determining risk-adjusted mortality rates with other procedures are currently not on the horizon.'

Finally, however much we wish it, advanced statistical modelling will not reveal factors that are otherwise obscure. When a clinician complains that a risk-adjustment process is inadequate, or does not correspond with clinical experience, the challenge is to find a way to enable the clinician to articulate his or her concern in such a way that it is open to measurement. Until that happens, the only reasonable assumption from the work to date is that severity of illness—at least as measured by clinical databases or laboratory results—does not account for all of the differences in death rates between hospitals.

2.7 Inter hospital variation and random variation

In Section 2.4.1, it was argued that V— variation in hospital mortality rates—would be made up from three components: V_Q = systematic variations in factors that influence mortality outcomes, $V_{C=}$ variations due to the clinical, demographic and casemix differences between patients present at the point of arrival in hospital, and V_R = random variations.

With methods for the computation of V_C established, the issue of random variation now has to be tackled.

If mortality is to be used as an indicator of safety and quality, then, like all indicators, it has to be reliable and valid. In psychometric practice, reliability is examined before validity. A reliable indicator may not be valid, but an unreliable indicator cannot be valid as its values cannot be interpreted.

The reliability and validity of indices of in-hospital mortality depend on the quality of measurement of relevant characteristics of hospital cases (e.g. number of diagnoses, vital status at the end of an episode of care).

2.7.1 Measurement

Although concerns have at times been expressed as to the accuracy of coding of diagnostic information within administrative data sets (Scott & Ward 2006), the extent of such disagreements in Australia at least are modest, and certainly appear to be no greater than found in the daily interactions between colleagues within the same team or discipline. Apart from diagnoses, the data elements in administrative data sets have generally been chosen because they are robust, straightforward to collect and enumerate and, in the case of the Australian National Hospital Morbidity Data collection, come with very explicit rules for their definition and tabulation. Coding audits constitute the test of inter-rater reliability relevant to assessing the utility of risk-adjusted measures of hospital mortality. Those audits commonly lead to no more than a small percentage of cases being re-coded: implying an acceptable level of inter-rater reliability (AIHW 2007).

It must be noted that although the fact of death will be accurately recorded, it is likely that there can be differences in relation to the proximate cause of death, as reported at death certification (Scott & Ward 2006). Fortunately, hospital mortality measures do not make use of the aetiological factors reported in death certificates, so that is not an issue of relevance.

2.7.2 Random and systematic variation

From Nightingale onwards, variations in hospital mortality rates have been taken to indicate variations in the safety and quality of the care provided. If hospital mortality rates are subject to large amounts of random variation, then they are outside the control of the staff in the hospital. Labelling a hospital as unsafe, when its results at any one time could vary between those considered safe and those considered unsafe solely due to chance, would be unreasonable for the staff and cause undue concern among current and potential patients. The reliability of the measure is clearly of great importance.

Random variation is present in the observation of all phenomena, though that is minimal in relation to the fact of death. The issue here is not the fact of death; it is variations in the observed death rate. Because patients are not randomly assigned to hospitals, the test that is applied to any one hospital is: 'does the observed mortality rate differ significantly from the rate that would have been expected if the patients had been treated in the 'average' hospital in the population of hospitals studied?'. Because inferences about hospitals are based on the size of the differences between the observed and expected mortality rates, the 'test-retest' question in relation to hospital mortality is whether the magnitude of differences remains similar when a hospital is studied again at a later time (assuming that the hospital's casemix did not change materially).

This question has been assessed in a number of ways: some more directly relevant than others.

A small group of studies in the 1990s (reviewed in Thomas & Hofer 1998) were conducted with the stated aim of examining the role randomness played in explaining hospital death rates (Zalkind & Eastaugh 1997: Thomas & Hofer 1999). Those studies all used broadly similar strategies, though they varied in scale and method. They all took as their starting point the assumption that variations in hospital mortality were a consequence of poor care, with poor care being identified via adherence to process measures. Then pre-existing external sources of information were used to specify the mortality implications of poor care, and these external parameters were then used to test the extent to which mortality outcomes in specific data sets could be attributed to poor quality. Simulation techniques were used to test the strength of relationships between mortality and poor quality, with Monte Carlo simulation being used to create multiple runs of the simulation equations under conditions of variation of the specified model parameters.

Because of the reliance of process measures as the measures used to infer poor quality, the studies were all on restricted ranges of diagnoses. The Zalkind (1997) study was entirely hypothetical, whereas Thomas and Hofer (1999) examined patients with CABG, AMI, stroke, pneumonia or congestive heart failure.

A careful examination of the analyses makes it clear that all those studies were in fact assessing the sensitivity and specificity of hospital mortality rates as indicators of hospital quality, with adherence to processes being the 'gold standard' against which hospital mortality was being assessed. Considered in that context, hospital mortality had low specificity in that there was a considerable risk that a hospital with a varying mortality rate might be flagged as low quality even though its quality, as measured by process adherence, was acceptable or high. Monte Carlo simulation showed that the poor performance in terms of this criterion was mainly the consequence of random variation.

Such studies are of interest, but they are of secondary importance here. They have cross-sectional designs rather than longitudinal, and so cannot measure variation over time in the absolute and relative performance of hospitals. As will be discussed later, the realisation of the extent and seriousness of adverse events during hospital care and the scale of their mortality outcomes, as crystallised in reports such as 'To err is human' (Kohn et al. 2000), and the 'Quality in Australian Health Care' study (Wilson et al. 1995), have altered the landscape in hospital mortality studies, and challenged pre-existing assumptions about how quality is identified and assessed.

Of more direct relevance are studies that look directly at test-retest or repeated measures issues. Marshall et al. (1998) described the development of a time series monitor of outcomes for patients undergoing CABG procedures in Veterans Affairs (VA) hospitals in the USA. Their concern was that monitoring the performance of VA hospitals by cross-sectional comparisons of performance would miss issues such as hospitals whose rates remained static although there was a general trend towards improvement, or hospitals whose results improved or deteriorated in a substantial way over time, despite the absolute mortality rates not being deviant enough to attract attention on cross-sectional study.

Implicit in such a design is the assumption that mortality rates are sufficiently predictable and stable over time that variations from the usual patterns will stand out. The study examined 11 six-month periods. The risk-adjusted mortality rates for patients undergoing CABG in 30 out of 43 hospitals were stable over the whole period, four hospitals had significantly high ratios over the whole period, and one significantly low. There was some movement in the rates for the remaining eight hospitals.

Birkmeyer and colleagues in the USA have been investigating the relationship between volumes of procedures performed, and subsequent mortality, for some time. As part of their work (Birkmeyer et al. 2006), they examined the value of historical mortality rates and procedure volume as predictors of subsequent performance on four high-risk surgical procedures (CABG, elective aortic aneurysm repair, oesophageal cancer resection, and pancreatic cancer resection).

They accessed the Medicare and Medicaid records for all patients undergoing these procedures over the period 1994 to 1997. Risk adjustment was undertaken for each procedure using the information in the Medicare data file: namely age, sex, race, admission status, socioeconomic status (defined as mean Social Security Income for the postcode of residence), and comorbidities aggregated into Charlson Index scores.

Morbidity rates in each hospital were then transformed using the t-statistic. The t-expected mortality is the difference between the observed and expected mortality, divided by the standard error of the expected mortality. This allows an adjustment for the variance due to small sample size, and tends to dampen the extreme mortality rates observed in hospitals with small case loads, moving them towards the mean. They then divided hospitals into quintiles of mortality for the period in question. Assignments to a quintile for the period 1994 to 1997, along with procedure volumes, were used to predict mortality during the subsequent two year period 1998–1999. Predictions were per procedure, and historical mortality predicted subsequent mortality for CABG, AAA and pancreatic resections, but not oesophagectomy. Historical mortality predicted 54% of subsequent mortality in CABG (compared with hospital volume, which only predicted 9%). It predicted 35% of mortality in AAA repair, and 41% in pancreatic resection. The same analysis was undertaken for the periods 1996–1999, and 2000–2001, with similar results.

Although not a conventional test-retest study, risk adjustment in these studies renders the populations similar to each other in relation to patient-level variability over time. The location of a hospital in a mortality quintile predicts its future location for the same procedure, implying that the measure—relative risk-adjusted hospital mortality—remains stable over time. If the differences between hospitals were solely a consequence of random variation, this would not be the case, and historical mortality could not predict subsequent mortality.

Finally, a study published recently has examined this directly (Heijink et al. 2008). Heijink and colleagues examined hospital mortality in all hospitals in Holland over the period 2003 to 2005. Risk adjustment was by means of age, sex, primary diagnosis, length of stay and admission status. The analysis was confined to those primary diagnoses causing 80% of hospital deaths. The HSMR was calculated for each hospital on the Dutch National Medical Registration. Nine of the 101 registered hospitals were excluded because of insufficient registration of separation data.

The aim of the study was to assess variation within hospitals over time and between hospitals in relation to a variety of organisational and environmental factors. Only the results in relation to variation within hospitals over time will be discussed at this point.

A two-level multi-level model was constructed to look at time trends. The results showed that there was a significant overall decrease in HSMR over the period in question, and that most of the variation in HSMR was caused by variation between hospitals rather than variation within hospitals over time.

Smith (Smith, 1994)—drawing on an earlier study (Smith et al. 1991) of Medicare data of over 41,000 patients in 81 hospitals and other studies—used complex statistical reasoning to partition the variance in hospital mortality into the three components described earlier (V_C = 50%, V_R = 15%, V_Q = 35%). Although no subsequent analyses have confirmed his partitioning, it is possible to draw some overall conclusions on the fact and partitioning of variability of hospital mortality.

First, it is clear that hospitals vary substantially in their mortality rates. Second, risk adjustment using the data elements in administrative data sets provides an acceptable level of discrimination in relation to hospital-level outcomes (though not, of course, prediction for individual cases). Third, after risk adjustment, the residual variations between hospitals have a substantial systematic element and the extent of random variation is not so great as to invalidate the use of hospital mortality as an indicator.

Whatever the factors that cause hospitals to differ, they tend to persist over time. Thus, in a recent publication from the Canadian Institute for Health Information (CIHI 2007), the HSMR outcomes from a large number of named Canadian Hospitals were computed and tabulated for the three 1–year periods from 2004–2005 to 2006–2007. The HSMRs and the confidence limits for the in-hospital mortality of each hospital were reported.

Using the simple expedient of saying that an HSMR that was above 100, accompanied by confidence limits that did not cross the 100, indicated a high-mortality hospital, and a hospital whose HSMR was below 100 and whose limits did not cross 100 indicated a low-mortality hospital (and all others were intermediate): 12 hospitals were low for each of 3 years, 36 were intermediate in each of 3 years, and 10 were high for the 3 years. Fourteen hospitals shifted between intermediate and low in one or more years, and twelve between intermediate and high. Only one hospital moved between all three levels in the 3–year period: and it went from being a high-mortality hospital to low, then intermediate.

Once it is agreed that variation is a fact, that it tends to persist after risk adjustment, and that in the absence of intervention it tends to remain stable over time, it becomes meaningful to examine to what the variation may be attributed.

2.8 The relationship between variations in hospital mortality and other measures

A modern general hospital is among the most complex of all human enterprises. Thousands of staff from a myriad of professional backgrounds—deploying varied and complex technologies, faced with patients whose combinations of principal and secondary diagnoses and other care needs are effectively infinite in number—make decisions whose implications are uncertain yet which can materially influence the very survival of the patients under their care. Is it not surprising then, that the literature on what it is that influences variations in hospital mortality rates is at times confusing and hard to follow. Some things, however, are fairly clear.

2.8.1 Structural characteristics

The structural characteristics of hospitals, including their staffing, their facilities, and possibly their role as teaching hospitals, are important but inconsistent predictors of in-hospital mortality (Silber et al. 1995).

For instance, Krakauer et al. (1992), in a broad-based study of mortality of Medicare patients treated in 84 hospitals across the USA, found that hospitals with a higher proportion of registered nurses or board-certified physicians, or with a greater level of access to high-technology equipment, had lower risk-adjusted mortality rates.

New York City has municipal public acute-care hospitals, and a large number of voluntary (private) hospitals. Shapiro et al. (1994) studied mortality rates for AMI, pneumonia, stroke, head trauma and hip repair in both municipal and voluntary hospitals. After risk adjustment using a wide range of secondary diagnoses, they found that there was increased mortality in the municipal hospitals for stroke and head trauma.

In an early study from New South Wales, Corben et al. (1994) looked at the variation in risk-adjusted mortality rates between different kinds of hospitals in New South Wales. The analysis showed that there were differences in mortality outcomes between hospital types (e.g. Principal Referral, District Hospitals) but the differences were not tested statistically.

Birkmeyer is a consultant to the Leapfrog Group in the USA, which promotes evidence-based purchasing amongst funders and purchasers of health care. In a series of large scale studies (e.g. Birkmeyer et al. 2002), Birkmeyer and colleagues have explored the relationship between volumes of surgical cases treated in hospitals, and hospital mortality, They demonstrated that, within the USA hospitals studied, there is a relationship between high volumes of certain high-risk cases treated and lower levels of hospital mortality.

In the study described previously, Jarman et al. (1999) looked at mortality rates for hospitals throughout England, and found that the best predictors of variations in hospital mortality were the numbers of hospital doctors per 100 beds and the numbers of general practitioners (GPs) per 100,000 population of the population served by hospitals.

In the recent large scale study of Dutch hospitals, Heijink et al. (2008) studied the relationship between variations in HSMR and a wide variety of structural characteristics of hospitals throughout Holland. The study used a sophisticated two-level multi-level random effects model to assess within hospital variation over time (previously discussed) and the influence of structural and input factors on inter-hospital variation. In the final analysis, factors such as socioeconomic status of the patients treated, numbers of nurses and doctors per bed and bed occupancy rate did not have an independent influence on mortality, though numbers of GPs per 10,000 occupants, and hospital type (teaching or non-teaching) did.

It is difficult to interpret the significance of the influence of community-based medical care in the Jarman and Heijink studies. Although it may be inferred that a relative lack of GPs might lead to patients who arrive at hospitals in a more severely ill state, that relationship has not been demonstrated empirically.

2.8.2 Performance, safety, quality

Performance

Since the publication of the HCFA studies in the 1980s, there has been a continuing interest in the search for measures of hospital performance. This has been fuelled by two major concerns. First, as health care has become increasingly expensive— particularly in the USA, but elsewhere also—but without clear evidence of the benefits of increased expenditure, efforts have been made to evaluate the performance of hospitals, to improve them; second, to provide guidance both to patients and to insurers or other purchasing groups, such as HMOs.

Safety

The *Compact Oxford English Dictionary* defines safe as 'protected from danger or risk; not causing or leading to harm or injury, and; (of a place) affording security or protection'. So, hospitals with relatively higher mortality rates are less safe overall than hospitals with lower mortality rates. That is self evident; when, after risk adjustment and allowing for random variation, mortality rates differ between hospitals, those hospitals with higher mortality rates afford their patients less security or protection than those with lower rates.

But this only applies at the hospital population level; and it is an increase in relative risk. It is quite inappropriate to deduce a conclusion for an individual on the basis of aggregate or population data: this is known as the ecological fallacy. Its force may be gathered from trying to deduce Sir Donald Bradman's batting average from the average for the Australian teams that he played in. The characteristics of a group may not be shared equally by all its members and, in population terms, the risks of the population at a whole are not equally shared by all its members. A patient with a particular risk profile may still be better off in a high-mortality hospital (Heijink et al. 2008), if that hospital is used to dealing with his or her condition. Furthermore, there is no a priori reason to assume that a hospital with a low-risk profile and a low-risk casemix would continue with that profile if it was faced with a higher risk case load.

Quality

What characterises quality in health care is not easy to pin down, and the relationship between in-hospital mortality and hospital quality measures is not clear. A dictionary definition of quality is that it is an essential or distinguishing characteristic. In common usage, the term tends to imply positive characteristics. What then, are the essential or distinguishing characteristics of high-quality health care?

Campbell et al. (2000) made a useful distinction between generic and disaggregated definitions of quality. A number of generic definitions of quality from fields outside health care base their definition on the viewpoint that a quality product or service is one that meets the requirements of those who use it. Thus a quality product or service is one that is fit for purpose or fit for use. Montgomery (2001), arguing from a statistical quality control viewpoint, defined quality as being inversely proportional to unwanted or harmful variability.

Within health care, the Institute of Medicine defined quality as the 'degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge' (Lohr 1991). Whatever the appeal of the generic definitions put forward by such bodies, they are hard to operationalise, and although disaggregating quality into a set of component parts emphases its complexity and multidimensional nature, the components are generally easier to measure.

A characteristic and much quoted multidimensional model is that of Maxwell (1984) who defined quality in relation to access to services, relevance to the needs of the whole community, effectiveness, equity, social acceptability, and efficiency and economy. That kind of multidimensional view is best understood in relation to a health service as a whole, rather than to an individual encounter within that service.

Donabedian has been the most influential voice in relation to quality at the level of the individual encounter. As he says in his landmark article 'Evaluating the quality of medical care' (Donabedian 1966: 163), he 'remained, by and large, in the familiar territory of care provided by physicians and has avoided incursions into other types of health care.'

Donabedian proposed that the quality of medical care be assessed in relation to three components—structure, process and outcome—of which structure has been dealt with above. Donabedian (1966: 186) recognised that outcomes validate other measures ('the validity of all other phenomena as indicators of quality depends, ultimately, on the relationship between these phenomena and the achievement of health and satisfaction') but introduced into common parlance the notion of measures of process as indicators of quality.

Process quality relates to an assessment of the interactions between clinicians and patients, and may be considered to have two elements (Schuster et al. 1998): technical process quality and care in relation to professional standards. Technical medical quality was simply described by Donabedian (1966) as 'whether what is known to be 'good' medical care has been applied'. By that he meant the skilful application of clinical care in the broad sense. It is clear that holistic assessments of that kind can only be made by judges who are themselves skilled: examining a range of information collected during an encounter. Such a strategy has come to be termed an implicit evaluation of care.

Broad-based evaluations can be distinguished from process quality as measured by process indicators. There, an assessment is made of the extent to which a specific process of care has been performed, defined either by reference to the scientific literature, or an expert panel, and deemed to represent appropriate care for a particular condition or set of circumstances. Most feasible process measures are usually indicators for a very specific element of the care process rather than comprehensive measures of how care is actually delivered (Rubin et al. 2001)—the hope being that the part is indicative of the larger whole.

It is the link between measured process and hospital mortality outcome that is most problematic. The underlying dilemma is clear.

Process measures provide direct feedback to professionals about measurable changes of practice; for example, ‘the percentage of eligible cases of patients with AMI who leave hospital with evidence-based treatments that will reduce the risk of recurrence’ is a measure that provides information that can be acted on.

But the link between specific process steps and overall hospital mortality is less clear because many of the factors that might affect mortality are outside the direct control of the practitioner. As Donabedian (1966: 181) puts it ‘Care can be good in many of its parts and be disastrously inadequate in the aggregate due to a vital error in one component’. Nevertheless, for the patient who is the subject of treatment, the process steps in his or her care are of little direct interest—what interests the patient is the outcome and, most interesting of all, the question of survival.

So is survival the gold standard of quality, and are measures that do not correlate with mortality poor measures of quality? Or is adherence to process standards the essence of quality, and measures that do not relate to variations in process adherence inappropriate measures of quality? Although this is clearly a matter of viewpoint, it is not simply a matter of semantics.

Take the following contrasting views. The Hospital Quality Alliance is a national public reporting program in the USA—initiated by the US Department of Health and Human Services—collecting data on a set of process measures for heart attack, heart failure, pneumonia and surgical site infection prevention. As Jha et al. (2007: 1105) point out, the indicators were developed with a broad consensus from experts, and from the research literature, but their performance ‘in the real world in identifying hospitals with better outcomes, such as lower risk-adjusted mortality across a number of clinical conditions is unknown’. Only if this relationship is confirmed can the measures be useful for quality improvement programs

However, in a review of a series of studies of the relationship between the Health Quality Alliance-supported process measures and hospital mortality, Shih and Schoenbaum (2007) found only a modest relationship between the measures and short-term mortality. As they say, equivocal results of this kind lead to criticism that such measures have only a limited value as tools for informing consumers about quality of care, or guiding payers seeking value in pay-for-performance programs (Horn 2006). Werner and Bradlow (2006) were similarly concerned that their findings of only a modest relationship between performance on process measures and risk-adjusted mortality rates—in a large scale study of Health Quality Alliance—supported process measures and mortality outcomes—would be inferred as meaning that the ability of performance measures to detect clinically meaningful differences across hospitals would be questionable.

A contrary view is exemplified in a recent comprehensive review of the relationship between quality of care and risk-adjusted mortality by Pitches et al. (2007: 1). The reviewers begin by partitioning mortality into patient casemix factors, random variation and a residual unexplained mortality (described as systematic variation above). The authors state that this unexplained component may ‘implicate quality of care’ and lead naturally to the ranking [in league tables] of hospitals with an implied correlation with quality of care. They go on to explicitly equate quality of care with adherence to existing evidence-based standards of clinical care, and seek to determine if hospitals with higher risk-adjusted mortality rates, provide poorer quality of care so defined. So, in this view, adherence to evidence-based standards of clinical care is the gold standard of quality against which mortality is assessed. This position has been re-stated particularly clearly by Shojania & Forester (2008: 153) who state that 'for the hospital standardised mortality ratio to represent a valid performance measure, it must correlate with accepted measures of quality'.

With these basic issues in mind, it is possible to begin to look for underlying patterns in the extensive literature that has accumulated in this area. This is a partial review only: more comprehensive analyses can be found in Iezzoni (1997a), Thomas and Hofer (1998), and Pitches et al. (2007).

Firstly, there are those studies that have gone from outcome back to process: that is, risk-adjusted mortality rates have been calculated, then processes within contrasting groups of hospitals have been examined. In an early study, Knaus et al. (1986) ranked intensive care units (ICUs) on mortality outcomes using the APACHE 11, and then undertook management audits of the units. The hospital with the lowest mortality ratio had a number of structural characteristics thought to be associated with good ICU care (e.g. 24 hour cover by a unit physician) and these were in contrast to the worst performing unit, where poor communication between the unit physician and the nursing staff was also noted. The small numbers and the very subjective aspects of the management audit make it hard to draw conclusions from this study.

The issue of small patient numbers is also found in a much quoted study by DuBois et al. (1987). In a rather complex design, they first created a crude risk adjustment that made no attempt to take comorbidities into account, and used that to rank hospitals in a provider-owned chain. They then studied six of the high-mortality outliers, and six of the low mortality outliers. Case records for a total of 378 patients with AMI, stroke or pneumonia were studied. A structured review against explicit criteria (generated by a panel of experts) was conducted by one of the researchers who was a physician. That physician also dictated case summaries for the 182 patients who died during their hospitalisations.

The extracted data was used for two purposes: firstly, a severity based analysis was conducted, allowing for more sensitive risk adjustment for each primary diagnosis. The performance against the explicit criteria was also reviewed and found not to vary between the high and low performing hospitals. The case summaries were then reviewed by external assessors, who looked at the overall care provided and rated the deaths as preventable or not preventable. After risk adjustment, the high-mortality hospitals were rated as having a greater proportion of preventable deaths for pneumonia and stroke, but not AMI.

The study is described in some detail because it reveals the complexity of the methods required to undertake an implicit review. Also noteworthy was that there was only modest inter-rater reliability between the assessors in relation to the outcome of implicit review.

Similar methods were then used in studies by Best and Cowper (1994), Goldman and Thomas (1994) and Gibbs et al. (2001). In each case, the potential preventability of deaths of patients who had died in hospitals with high (Goldman and Thomas 1994) or high and low mortality rates, (Gibbs et al. 2001) was assessed by independent assessors. Although in both cases higher overall mortality was associated with deaths that were deemed more likely to have been preventable, the associations were generally modest.

Park et al. (1990) in a RAND Corporation study, used HCFA data to identify high outlier hospitals, and compared a representative sample of over 2000 patients with either congestive cardiac failure or AMI. Quality of care was examined by a detailed case note review, in which quality of care was assessed in relation to an explicit set of processes, though the processes assessed were quite broad, and included physician and nurse examination, diagnostic tests and use of therapeutic and intensive services. Although, at the patient level,

higher severity and poorer quality of care were associated with higher mortality, no hospital-level effect could be detected
(a demonstration of the ecological fallacy). Interestingly, simulation was used to assess the extent to which variations in hospital mortality could be attributed to random variation. Although that proportion was substantial, the non-random variation was statistically significant and clinically important.

A quite different approach that made it possible to overcome the problems of small sample sizes, but traded size for credibility, made use of the fact that files of patients for whom USA hospitals claimed re-imbursements were independently assessed by peer review organisations in each state (Hartz et al. 1993). The Peer Review Organisations review about one in four records. Nurse reviewers look for a specified set of quality-of-care performance problems (quite diverse and widely drawn) and, once a problem has been identified, a physician review confirms the problem or not.

Although there were modest, but statistically significant, correlations between problem rates as specified by the Peer Review Organisations and risk-adjusted hospital mortality rates, at the state level, there were major differences between the Peer Review Organisations in each state. Hence, the findings of Hartz et al. (1993) and Thomas et al. (1993), which also used Peer Review Organisation assessments, are hard to interpret.

Finally, there are a number of other studies (reviewed in detail by Pitches et al. (2007)) that use explicit review to assess compliance with process measures for one or more specific conditions in patients treated in groups of hospitals, and assess the association of process measure compliance with risk-adjusted hospital mortality for those conditions. The outcomes of these studies are in line with the outcomes of the Health Quality Alliance process measures.

The literature reviewed in this section demonstrates that the relationship between process measures and mortality outcomes is inconsistent. Further work in this area should continue to be monitored.

2.8.3 Studies aimed at changing hospital mortality rates

The recognition in recent years of the pervasive nature of adverse events during hospital care, and their mortality and morbidity implications, has begun to change the context of discussions about hospital mortality.

As the previous section demonstrates, for many years the concentration was on hospital mortality as an indicator of quality, when quality was associated with the performance of clinical practitioners in relation to what might broadly be termed evidence-based care. Do practitioners do what is thought to be necessary, or at least practise in conformity with the best evidence for what ought to be done? In that context, a good quality hospital is one that provides the right care. But as Donabedian (1966: 182) points out, the relation between structure, process and outcomes is not simple:

> ‘In healthcare, each event is an end to the one that comes before it and a necessary condition to the one that follows. This indicates that the means–ends relationship between each adjacent pair requires validation in any chain of hypothetical or real events. This is … a laborious process. More commonly… the intervening links are ignored. The result is that causal inferences become attenuated in proportion to the distance separating the two events on the chain.’

There are very many steps between a specific process measure (giving aspirin on arrival in hospital for patients with AMI) and overall in-hospital mortality. And studies are now emerging that describe hospitals' efforts to reduce overall mortality rates directly, rather than looking solely at specific process steps.

In 2000, the Walsall Hospital NHS trust had a HSMR of 130: the highest of all acute hospitals in England. In response, seven clinical governance groups were formed to implement changes across the whole range of clinical disease areas, together with a wide variety of management areas including bed management, information services, discharge liaison, integrated care pathway development, and many others. By the end of 2004, the HSMR had dropped to 92.8 (Jarman et al. 2005).

It could be argued that what was accomplished here was no more than statistical regression to the mean, or a more causal effect resulting from public scrutiny causing a poorly performing hospital to get back into line (akin to the 'Hawthorne effect'). Any change, no matter what its impetus, would have had the same outcome.

The case study of the Bradford Teaching Hospitals Trust (Wright et al. 2006) is particularly interesting in this light. The Trust is a large (1200 bed) acute service which, in 2002, was a low mortality Trust in terms of HSMR. Nevertheless, in 2002 the Trust chose to focus on hospital mortality, with a commitment to eliminate all unnecessary hospital deaths. The program of work that followed was very diverse. Following a review of a consecutive series of hospital deaths, a high prevalence of sub-optimal clinical observations, medication errors and hospital infections was noted amongst the patients that died. A wide variety of corrective actions were initiated in all relevant areas. Also, a monthly monitoring program for hospital deaths using a statistical control chart for hospital mortality was begun. (Statistical control charts are discussed further on in this report.)

In Bradford, the effect of the mortality reduction program was to significantly reduce the hospital HSMR from 94.6 at the start of the program to 77.5 three years later. The Bradford Trust began its work after enrolling in an Institute of Healthcare program: Improvement Partnership for Hospitals. Gilligan and Walters (2008) described the experiences of the East Lancashire Hospitals Trust (Royal Blackburn Hospital) following enrolment in that same program. Their focus became improving the flow of patients through the hospital by a combination of activities including changes to medication charts and physician rounds, redistribution of bed stock and the introduction of a critical-care outreach service, plus intensive monitoring of outcomes using control charts. The Trust was never a high-mortality outlier—though its HSMR was above the national average—but over the period of the study, the HSMR declined substantially.

The large-scale 100,000 Lives Campaign initiated by the Institute for Healthcare Improvement is aimed directly at hospital mortality: reducing mortality by a series of broad based improvement strategies, rather than through the medium of adherence to narrow-focused process measures for specific conditions. The strategy is not without its critics (Auerbach et al. 2007), but is defended indirectly in Berwick (2008).

2.8.4 Summary

The performance of hospitals will continue to be scrutinised, and measures will continue to be devised to open up the historically rather hidden world of hospital outcomes to external inspection. Mortality is one such measure, and although its status as a measure of safety seems secure, its role in specifying hospital quality is subject to the difficulties and ambiguities inherent in the concept of quality.

Feasible and reliably measurable process measures tend to be of very specific elements of care, and there are likely to be very many unmeasured steps between such process elements and the survival or death of a group of patients—as the mortality reduction programs described above infer. Furthermore, there is a logical fallacy at the heart of referring back from mortality to quality when (and if) quality is defined in relation to performance levels on a set of specific process measures. If quality is synonymous with p, and p→m (mortality), that does not necessarily mean that p←m, because m is not identical to p. All cherries are red, and all cherries are fruit, but this does not mean that all red fruit are cherries.

Another way of looking at this is the fallacy of composition. The fallacy of composition is committed when a conclusion is drawn about a whole, based on the features of its constituent parts, when there is no justification for drawing this inference. For example, every player on the team is a superstar, so the team is a great team. This is not necessarily so, because the superstars will not necessarily play together well, and so could form a very weak team. Teamwork is a quality of interaction and not a matter of simple addition. Similarly, in a hospital, individual staff may perform specific process measures with great accuracy, but modern health care depends as much on teamwork as individual competence, and the health-care team as a whole may perform poorly (Lemieux-Charles & McGuire 2006), and so increase mortality risk, despite the team being made up from conscientious and caring practitioners.

2.9 Presentation of information about in-hospital mortality

2.9.1 Methods of presentation

Variations in hospital mortality rates are analysed and disseminated in an effort to influence the recipients to look further at health-care practice. Thus the mode of presentation of the results of analyses of mortality rates is of considerable interest.

Goldstein and Spiegelhalter (1996) have provided a review of issues in this area. Drawing on examples from education and health care, they make the important point that comparisons must take context into account. Risk takes account of patient characteristics at entry into a hospital, in the same way that comparisons of schools performances should take account of the status of the children on arrival at a school. Goldstein and Spiegelhalter (1996) further argue that the need to contextualise does not stop at the institutional level, but needs to be considered at state and national levels.

So, accepting that simple comparisons of mortality rates that are not risk adjusted will almost always be confusing, there are only a small number of practical alternatives for presenting the information.

As previously described, risk adjustment of hospital mortality always involves a comparison between observed and expected mortality rates for a set of institutions or services. Although a number of different ways of generating indices from this comparison have emerged over the years (starting, in Australia, with an interesting early paper by Duckett & Kristofferson

1978), the HSMR has emerged as the standard in this area, and so is the index of hospital mortality discussed further.

Institutional HSMRs can simply be listed. But any single HSMR needs to be accompanied by a measure of the uncertainty of the value. The conventional method of doing that in scientific biomedical practice is to calculate the confidence intervals around each HSMR, usually using 95% confidence limits (e.g. CIHI 2007). The 95% confidence limits represent the range within which a particular parameter will be found 95% of the time on repeat testing of a population, so the width of the confidence limit gives an indication of the uncertainty or precision of the parameter. Wide confidence limits commonly occur when sample sizes are small.

The Canadian National Study of HSMRs, referred to above, simply listed each participating institution, together with its HSMR and the confidence limits. It made no explicit inter-institutional comparisons: leaving that to the reader.

League tables

League tables in which hospitals are ranked in relation to their particular HSMRs, are an explicit method of providing inter-institutional comparisons. The Dr Foster group in the UK has provided several non-peer-reviewed reports in which hospital are ranked according to their HSMRs.

Typically, most hospitals in a country have HSMR values that are quite close to one another, especially after adjustment for casemix. League tables tend to encourage unwarranted emphasis on small and unimportant differences in the rates, because they can translate into large differences in the ranking of hospitals with similar rates.

League tables are improved by the addition of confidence intervals. But no matter how much effort is put into explaining uncertainty and variation, it is hard not to assume that being 24th in a table of institutions ranked from 1 to 100 really means that the institution in question is superior to the 25th institution, and much superior to the 35th, even if all of those institutions have overlapping confidence intervals and cannot be said to differ significantly. So, whatever their attraction, from a statistical and epidemiological viewpoint the presentation of HSMRs in a simple league table format is hard to support.

Caterpillar plots

Another method is to present HSMRs (Goldstein and Spiegelhalter 1996) in the form colloquially known as ‘caterpillar plots’, in which the HSMRs and their confidence intervals are represented as a graphical plot, with individual institutions ranked by HSMR along the x-axis and the HSMR values shown on the y-axis.

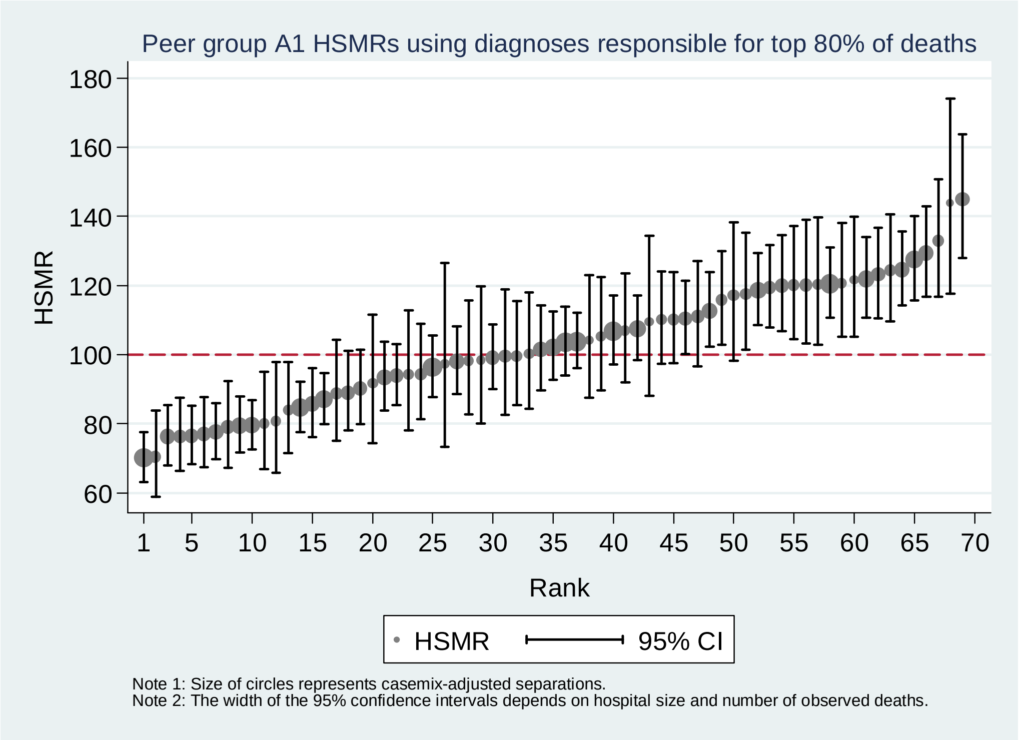
The two caterpillar plots below (Figure 2 and 3) are included here to illustrate this type of presentation. Each of the plots summarises data for one peer group of hospitals. The analysis underlying these plots has applied the same adjustment model to both peer groups (rather than analysing each group separately). Hence, it is computationally valid to compare the HSMR values in each of these plots. The values for hospitals in the A1[2] group are spread fairly equally above and below 100 (Figure 2). In contrast, the values for B1 hospitals are mainly below 100 (Figure 3). However, the interpretation of this difference between peer groups is complicated by their different casemixes. Adjustment for casemix based on data available in administrative data allows for part, but not all, of the difference. An apparently

[2] See Table 2 for information on the types of hospitals included in the peer groups.

low-risk group of hospitals will only be low risk for their casemix, not the casemix of larger hospitals.

The extent to which low-risk populations, as well as low-risk hospitals, provide an important opportunity for analysis is discussed in Coory and Scott (2007).

The examples of caterpillar plots presented in this section are typical of those in the literature. There may be potential to improve the performance of this type of plot as a graphical method to convey information about in-hospital mortality. We present and discuss some variations in Appendix 4.



Note 1: Size of circles represents casemix-adjusted separations.
Note 2: The width of the 95% confidence intervals depends on hospital size and number of observed deaths.

Figure 2: Caterpillar plot of variation in point estimates in HSMR for peer group A1

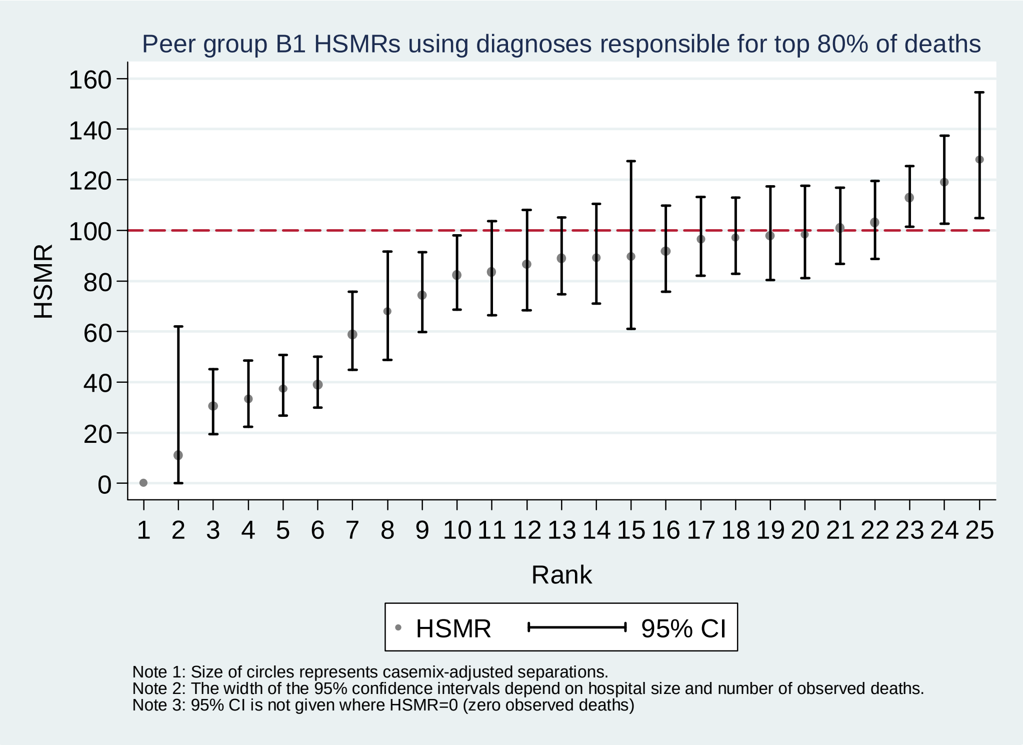


Figure 3: Caterpillar plot of variation in point estimates in HSMR for peer group B1

The obvious question is 'when is a difference between institutions important?' When the lower confidence limit of the estimate for any an institution is above the population average of 100, or the upper confidence limit is below 100, then that institution differs statistically from the population average. When a HSMR is so deviant that the institution not only fulfils the above criterion but is say 15% above or below the average, some analysts would declare the institution to be an outlier. Some would set even more rigorous criteria against which to assess outlier status and some would not set an outlier standard at all, but would just identify institutions at extremes. There is no absolute standard here.

It is also the case that when the confidence intervals of two institutions do not overlap, they are deemed to differ to a statistically significant extent from each other, and that is helpful when undertaking inter-institutional comparisons for sub-samples of institutions that appear at very low (or very high) risk overall—at least in relation to HSMR.

The results of the analysis of the Australian data sets are presented later in the form of a series of caterpillar plots, and their utility can be gauged from those presentations.

Funnel plots

A relatively recent innovation in the area of the analysis and presentation of HSMRs and other hospital performance indicators is the use of funnel plots, which were extensively developed by the Medical Research Council Biostatistics Unit in Cambridge in the UK (Spiegelhalter 2002: 2005) and are now coming to be seen as potentially preferable to caterpillar plots (e.g. Mohammed & Deeks 2008).

In the context of institutional comparisons, a funnel plot is an extension of a Shewart chart, or a statistical control chart. It is a method for detecting when a particular institutional outcome on a parameter, such as the HSMR, is so extreme as to constitute a potential case of 'special-cause' variation. This means that the variation is so great that it is outside the bounds of the underlying, or common, cause variability that is present in the usual outcomes of the parameter in question. When a control chart 'signals' special-cause variation, an investigation into potential causes should follow.

The method of computation of funnel plots is quite complex, although the results are presented in an easy to assimilate graphic. A relatively straightforward description is provided by the Dr Foster group in a recent non-peer reviewed (Dr Foster Intelligence 2007).

> 'Funnel plots (or control charts) are a graphical method used to assess variation in data and are used to compare different trusts over a single time period. These plots (HSMR funnel plots) show the position of each trust's HSMR. Control limits form a 'funnel' around the benchmark and reflect the expected variation in the data.
> [The wide base of the funnel demonstrates that as the number of separations involved fall, the size of the expected variation increases because the measure is less precise].

Each chart has five lines:

- a centre line, drawn at the mean (the national average RR=100)
- an upper warning line (upper 95% control limit)
- an upper control limit (drawn three standard deviations above the centre line-upper 99.8% control limit)
- a lower warning line (lower 95% control limit)
- a lower control limit (drawn three standard deviations below the centre line–lower 99.8% control limit).

Data points falling within the control limits are consistent with random or chance variation and are said to display "common-cause" variation. For data points falling outside the control limits, chance is an unlikely explanation, and hence they are said to display "special-cause" variation.'

Further discussion of methods of presentation is delayed until after the results of the Australian study are provided.

2.9.2 Public or private dissemination of mortality outcomes

There has been lengthy discussion over the years as to the legitimacy of public reporting of mortality data, in comparison to dissemination solely to the institutions themselves. The issue will not be discussed at length here for several reasons.

First, public dissemination of performance indicators is an area that has been comprehensive reviewed on a number of occasions, and there is little to add to recent reviews (e.g. Fung et al. 2008; Hibbard et al. 2005).

Second, the overall results are fairly clear. Public reporting has, at best, a modest impact on the public at large, but it has a more definite impact on providers of care: tending to increase improvement activities of a variety of kinds. It is not without its hazards however (Hibbard et al. 2005).

Third, public reporting of mortality, as well as many other outcomes, is already so widespread as to be the norm in the USA, the United Kingdom and Canada, and in the UK will become increasingly so if the reforms recently advocated by Lord Darzi are enacted. In Australia, the Queensland Measured Quality reports, first produced in 2004 (Queensland Health 2004) provide very detailed mortality and other information about the hospitals in Queensland, and the reports have been elaborated in various ways since then.

Finally, there will always be a necessary tension between the desire of establishments to maintain a good reputation and a public right to know. Media reports based on publicly available information have not always presented a completely accurate, or necessarily fair, representation of institutional or even personal outcomes. Public reporting does, however, guard against a tendency to withdraw support from analyses that may be seen as a source of embarrassment or distress—even if they are accurate—but it also places an obligation on the reporter to stringently guard against bias and misrepresentation.

2.9.3 Future developments of note

As well as providing a review of existing work, we have also been asked to comment briefly on any noteworthy trends in data gathering or analyses. We would say that the two most promising developments that will be implemented in the near future are the decision to require national coding of 'present on admission' indicators for all secondary conditions in the Australian National Hospital Morbidity set, and the wider application of data linkage. Some methodological developments also hold promise.

Present on admission indicators

One of the challenges for risk adjustment of performance indicators is a health-care version of the moral hazard problem. Coders are required to code complicating or comorbid conditions, irrespective of whether they were present at admission or occurred after admission. Some of those secondary conditions may have been the result of problems that occurred as a result of sub-optimal care. To risk adjust for them is to provide an allowance for poor-quality care rather than to reveal it by comparison of outcomes. For example, patient X had a presenting problem of severity Y, and was at low risk of death; having had a series of falls and a surgical site infection, he is now ranked as a high-risk patient, and his death is partially discounted for that reason.

One way to capture this in hospital data is to attempt to record which conditions were present on admission. 'Present on admission' codes require a coder to judge whether a secondary condition was, or was not, present on admission, and are mandatory for Australian public hospital-coded separation data from the beginning of financial year 2008–09. Present on admission coding has been practised for some years in California, and a recent study by Glance et al. (2008) demonstrates that present on admission coding is likely to considerably enhance the precision of mortality measures. Present on admission coding (known as C-codes in Victoria) has been in place in Victoria for some time, and a study by Ehsani et al. (2006) has shown that it is similarly informative there.

Data linkage

A second useful development is the increasing availability of data linkage. Two forms of linkage are relevant. The first is linkage within hospital morbidity data. Some people—particularly those with serious and persisting conditions—are likely to experience more than one episode of in-hospital care within the period covered by a study of in-hospital morbidity. Hospital inpatient administrative data files have generally been organised in a way that includes a record for each of these 'separations', but does not provide a good basis for grouping together the set of records referring to a particular person or reason for admission. Without this form of linkage, it is not possible to be sure whether a person whose episode of hospital care ended with transfer to another hospital, or with a 'statistical type change', died during the next episode of inpatient care. Even a person who separates with discharge home might have been re-admitted soon after, with the possibility of fatal outcome of that episode.

The second role of data linkage relevant to this type of work is linkage between hospital records and death registers (or the National Death Index). This is necessary to enable studies that include deaths soon after discharge.

Health data linkage systems also have potential to be used to assess individual health status over time. Such information might be found to improve risk adjustment.

Developments that enable such linkage are well-established in some parts of Australia (notably Western Australia and New South Wales) and are being put in place elsewhere (e.g. South Australia), but there is not yet a routine capability to enable the necessary linkage at national level.

Analytical methods

From a methodological point of view, the issue of the development of Bayesian regression models for use in large scale mortality studies (e.g. Austin 2008) is of interest, but will require further study. The approach has potential for analysis of smaller hospitals. Bayesian techniques have a number of adherents in the field of performance measurement, but the techniques can be complex and are not without controversy, and will require quite detailed assessment and testing before their strengths and weaknesses can be assessed in this context (Paul Aylin personal communication, 2008). Nevertheless, this approach is sufficiently promising to warrant exploratory use and further development.

Further developments in statistical process control methods for immediate monitoring of mortality and other performance measures are also clearly an area of great interest however (Duckett et al. 2007)

2.10 Conclusions

In 2006, Scobie et al. (2006)— drawing on the work of the National Performance Committee (NHPC 2004)—provided a useful set of criteria against which to assess the potential utility of a candidate health performance indicator. Those criteria can be used to assess variations in hospital mortality as a candidate indicator of hospital performance.

Scobie et al. (2006) stated that an indicator should:

1. **Be worth measuring—it should represent an important and salient aspect of the performance of the health system**. It is hard to argue against variations in hospital mortality on those grounds.
2. **Be measurable for diverse populations—the measure should be valid and reliable for general populations and the diverse populations in Australia.** Variations in hospital mortality rates are relevant to all populations studied, and are reliably reported.

3. **Be understood by people who need to act.** The fact of variations in mortality is readily and immediately understood. The remedial actions are less clear.
4. **Galvanise action—The indicators are of a nature that action can be taken at the national, state, local or community level by diverse groups of individuals.** Once the fact of variations in mortality are acknowledged, then actions take on some urgency, though, again, this is at an early stage and the necessary roles of the various levels in the health system are not yet clear.
5. **Be relevant to policy and practice.** Although the policies and practices that will directly focus on mortality reduction are yet to be generally agreed, the speed with which institutions have taken up the creation of medical emergency teams as a mortality reduction measure indicates that remedial actions can be developed and implemented on a wide scale.
6. **Reflect results of actions when measured over time.** The studies described earlier demonstrate that.
7. **Be feasible to collect over time.** This is clearly possible.

Variations in hospital mortality appear to fulfil all the necessary criteria to qualify as a performance measure. The more pressing question is 'how they should be used?'

The uncertainty surrounding the relationship between variations in hospital mortality and other measures of hospital structure and process mean that, in our view, variations in hospital mortality should be viewed as screening tools, rather than being assumed to be definitively diagnostic of poor quality. A screening tool is a signalling device. It is intended to signal that a problem may exist and that further detailed investigation is required.

With a screening tool, some lack of precision is accepted, because being too cautious in sounding a warning risks ignoring a problem in its early stages, when it may be more open to change. So, because of the uncertainty in the interpretation of mortality rates, it is inappropriate to use variations in hospital mortality to assert with confidence that a high-mortality hospital provides poor-quality care. That is a premature rush to judgment. High relative mortality should be seen as a prompt to further detailed investigation. The issues were well summed up by Donabedian (1966: 196). 'A final comment concerns the frame of mind with which studies of quality are approached. The social imperatives that give rise to assessments of quality have already been referred to. Often associated with these is the zeal of the social reformer. Greater neutrality and detachment are needed in studies of quality. More often one needs to ask "What goes on here?" rather than "What is wrong; and how can it be made better?" '