Performance Measurement for Health System Improvement

Experiences, Challenges and Prospects

Peter C. Smith, Elias Mossialos, Irene Papanicolas and Sheila Leatherman

HEALTH ECONOMICS, POLICY AND MANAGEMENT

European Observatory on Health Systems and Policies

WORLD HEALTH ORGANIZATION EUROPE

CAMBRIDGE

PART III

Analytical methodology for performance measurement

3.1 *Risk adjustment for performance measurement*

LISA I. IEZZONI

Introduction

Risk adjustment within health care aims to account for differences in the mix of important patient attributes across health plans, hospitals, individual practitioners or other groupings of interest before comparing how their patients fare (Box 3.1.1).

Box 3.1.1 Definition of risk adjustment

This statistical tool allows data to be modified to control for variations in patient populations. For example, risk adjustment could be used to ensure a fair comparison of the performance of two providers: one whose caseload consists mainly of elderly patients with multiple chronic conditions and another who treats a patient population with a less severe case mix. Risk adjustment makes it possible to take these differences into account when resource use and health outcomes are compared.

Source: Institute of Medicine 2006.

This straightforward purpose belies the complexity of devising clinically credible and widely accepted risk adjustment methods, especially when resulting performance measures might be reported publicly or used to determine payments. Controversies about risk adjustment reach back to the mid-nineteenth century. Florence Nightingale (1863) was criticized for publishing figures that showed higher death rates at London hospitals than at provincial facilities: 'Any comparison which ignores the difference between the apple-cheeked farm-labourers who seek relief at Stoke Pogis [sic] (probably for rheumatism and sore legs), and the wizzened [sic], red-herring-like mechanics of Soho or Southwark, who come into a London Hospital, is fallacious' (Anonymous 1864 pp.187–8). Other critics noted that many provincial

hospitals explicitly refused patients with phthisis (consumption), fevers or who were 'dead or dying', whereas urban facilities took everyone (Bristowe & Holmes 1864). Had the figures Nightingale published 'really overlooked the differences in relative severity of cases admitted into ... different classes of Hospitals ...?'[1] (Bristowe 1864 p.492).

Similar complaints echo 150 years later – risk adjustment methods are inadequate and failures of risk adjustment might affect the willingness of health-care institutions and practitioners to accept difficult cases and publicly release performance data. Certainly, there have been advances in what some consider 'the Holy Grail of health services research over the past 30 years' (McMahon et al. 2007 p.234). Statistical techniques for adjusting for risks are increasingly sophisticated. Reasonably well-accepted methods for capturing and modelling patients' clinical risk factors now exist for a variety of conditions, especially those involving surgery and risks of imminent death or post-operative complications. This brief chapter cannot hope to review the full (and growing) range of current risk adjustment methods which span practice settings from intensive inpatient to home-based care. Nevertheless, much remains to be done. In 2006, the Institute of Medicine (2006 p.114) highlighted the need for continuing applied research to support performance measurement, specifically calling for studies of risk adjustment methods. Commenting about inadequate performance measurement methodologies generally, it warned, 'data can be misleading, potentially threatening providers' reputations and falsely portraying the quality of care provided.'

This chapter explores basic issues relating to risk adjustment for quality performance measurement. Another important use of risk adjustment methods involves setting payment levels for health-care services. In 1983, Medicare introduced the earliest widely implemented risk adjustment method by adopting DRGs for prospective hospital payment. These are now utilized worldwide, albeit with nation-specific variations, especially throughout Europe. Langenbrunner et al. (2005) describe the various applications of DRGs for setting hospital payments. Hospital cases are assigned to pre-set reimbursement levels

[1] Nightingale (1863) used hospital mortality figures calculated by William Farr. This physician and prominent social reformer shared her passion for motivating hospital improvement through statistical analysis and comparing outcomes across facilities. Farr had conducted analyses for the Registrar-General since 1838.

(or relative weights) based primarily on patients' principal diagnosis, surgery or invasive procedure and whether they have significant co-morbidities or complications. DRGs have evolved over time, mainly to keep abreast of technological advances and newly emerging health conditions but also more recently to account better for severity of illness (US Department of Health and Human Services 2007). Other risk adjustment methods are used to set payment levels for capitated health plans, nursing home stays, home health-care episodes and other types of services. Risk adjustment for payment purposes raises special issues. In particular, critics worry that inadequate risk adjustment exacerbates incentives to avoid or limit care for very sick patients.

Cost-focused and quality performance-targeted risk adjustment methods share important conceptual foundations but are intended to predict different outcomes. Generally they have different specifications and weighting for risk factors but some aspects may overlap. In 2005 the United States Congress mandated that after 1 October 2008 Medicare would no longer pay hospitals for treating preventable complications that shift cases into higher-paying DRGs (Rosenthal 2007). The eight selected complications[2] are generally avoidable so this policy aims to stop financial rewards for substandard care. Pay for performance is another area where cost- and quality-focused risk adjustment may overlap (or collide). As described below, concerns about the validity of these measures (including the adequacy of risk adjustment) have taken centre stage in debates about these efforts worldwide.

Why risk adjust?

Rationale for risk adjustment

Health plans, hospitals, general practitioner practices or other health-care providers are not selected randomly. Many factors affect the way people link with their sources of care, including the nature of their health needs (e.g. acuity and severity of illness); financial resources; geography; previous health-care experiences; and their preferences,

[2] Medicare will not pay for any of the following acquired after admission to hospital: air embolism; blood incompatibility; catheter-associated urinary tract infection; pressure ulcer; object left in patient during surgery; vascular catheter-associated infection; mediastinitis after coronary artery bypass grafting; fall from bed (Rosenthal 2007).

values and expectations of health services. Not surprisingly, there may be wide variations in the mix of persons covered by different health plans, hospitals, general practioner practices or other health-care providers. These differences can have consequences. For example, older persons with multiple chronic conditions require more health services than younger healthier people and are thus more costly and complicated to treat. Most importantly from a quality measurement perspective, persons with complex illnesses, multiple coexisting conditions or other significant risk factors are more likely to do poorly than healthier individuals, even with the best possible care.

Most quality performance measures reflect contributions from various patient-related and non-patient factors. For example, hospital mortality rates after open heart surgery reflect not only the technical skills of the surgical team and post-operative nursing care but also the severity of patients' cardiovascular disease, extent of co-morbid illness and level of functional impairment. Screening mammography rates reflect not only recommendations from clinicians and the availability of the test but also women's motivation, ability and willingness to attend. Thus, a complex mix of factors contributes to how patients do and what services they receive. Patient outcomes represent a particularly complicated function of multiple interacting factors:

> Patient outcomes = *f* (effectiveness of care or therapeutic intervention, quality of care, patient attributes or risk factors affecting response to care, random chance)

Risk adjustment aims to account for the effects of differences when comparing outcomes across groups of patients. It assists in disentangling the variation in patient outcomes attributable to intrinsic patient factors (generally not under the control of clinicians or other health-care providers) from factors under clinicians' or providers' control, such as quality of care. Generally, it is critical to use risk adjustment before using patient outcomes to draw inferences about the relative quality of care across health plans, hospitals, individual practitioners or other units of interest. Risk adjustment aims to give outcome-based performance measures, what Donabedian (1980 p.103) calls 'attributional validity' – the conviction that observed outcome differences causally relate directly to quality of care rather than to other contributing factors.

Despite this straightforward rationale, critics warn that it may be quixotic to believe that quality of care variations can be adequately isolated by adjusting comparisons of patients' risks and other factors (Lilford et al. 2004 p.1147). As Terris and Aron (2009) observe in this volume, proving attribution may require exploration of causality from multiple and varied perspectives. Thus, risk adjustment performed in isolation can produce a false sense that residual differences among providers reflect variations in quality. Different risk adjustment methods can paint divergent pictures of provider performance according to their data sources, variable specifications and weighting schemes. For instance, different risk adjustment methods produced varying impressions of rankings of hospitals based on their relative mortality rates (Iezzoni 1997). Hospitals ranked highly by one risk adjuster may plummet in the rankings of another. Lilford et al. (2004 p.1148) note that, 'case-mix [i.e. risk] adjustment can lead to the erroneous conclusion that an unbiased comparison between providers follows. We term this the case-mix fallacy.' Nonetheless, without any risk adjustment, patient factors can hopelessly confound comparisons of outcomes and of other performance measures.

Consequences of failing to risk adjust

There can be serious consequences from failing to risk adjust before comparing how patients do across health plans or providers. Most importantly, the resulting information could be inaccurate or misleading and consumers, policy-makers and other health-care stakeholders will not have valid information for decision-making (Institute of Medicine 2006).

Intended audiences may grow to distrust, disregard or dismiss poorly-adjusted data. This happened after Medicare first published hospital mortality rates more than twenty years ago (Box 3.1.2). A 2005 national survey of American general internists found that 36% strongly agreed and 52% somewhat agreed that, 'at present, measures of quality are not adequately adjusted for patients' medical conditions.' Interestingly, 38% strongly agreed and 47% somewhat agreed that current quality measures 'are not adequately adjusted for patients' socioeconomic status' (Casalino et al. 2007 p.494). Without clinician buy-in, initiatives that use performance measures to try to

Box 3.1.2 Inadequate risk adjustment

In March 1986, the Medicare agency in the United States publicly released for the first time hospital mortality rates for its beneficiaries. According to governmental predictions, 142 hospitals had significantly higher death rates than predicted, while 127 had significantly lower rates. At the facility with the most aberrant death rate 87.6% of Medicare patients had died, compared with a predicted 22.5%. This facility was a hospice caring for terminally ill patients. The government's risk adjustment model had not accounted adequately for patients' risks of death.

Source: Brinkley 1986.

influence clinical practices will likely fail or confront controversy and challenges.

Pay-for-performance programmes are a case in point, usually at the forefront of risk adjustment debates. These initiatives aim to align payment incentives with motivations to improve health-care quality but many observers have raised another troubling possibility. If pay-for-performance measures are perceived as unfair or invalid because they do not account adequately for patients' risk factors, then clinicians or health-care facilities may game the system by avoiding high-risk patients who are unlikely to do well (Birkmeyer et al. 2006). To maximize the fairness of pay-for-performance measures, risk adjustment may need to consider not only patients' clinical characteristics but also their socio-demographic complexity and other factors that might affect adherence to treatment regimens, as well as screening and preventive care (Forrest et al. 2006).

Some observers worry that pay-for-performance incentives could potentially precipitate adverse selection – the pressure to avoid severely ill or clinically challenging patients (Petersen et al. 2006; Scott 2007). In addition, vulnerable subpopulations could lose access to care e.g. those with lower socio-economic status and a heavy burden of disease who tend to cluster in specific locales (e.g. distressed inner-city neighbourhoods): '... What happens to providers with a disproportionate number of high-risk patients? They can dump their patients, they can get paid less, or they can move' (McMahon et al. 2007 p.235).

This concern is bolstered by early experiences from the United Kingdom's NHS pay-for-performance initiative targeting general practitioners that began in 2004 (Roland 2004; Velasco-Garrido et al. 2005). Given the nature of some NHS performance measures (see below), general practitioners could perform better by excluding certain high-risk patients from reporting (Doran et al. 2006). Practices could game the incentive system by avoiding such patients or reporting that these patients were exceptions to required clinical actions or outcomes. Evidence of widespread gaming has failed to materialize but a small minority of practices (91 or 1.1%) excluded more than 15% of their patients from performance reporting (Doran et al. 2006). In countries like New Zealand that have not yet widely implemented pay for performance, NHS experiences raise fears of potential gaming incentives and other unintended consequences, leading to caution in specifying initial performance measurement sets. The Effective Practice, Informatics and Quality Improvement (EPIQ) programme at the University of Auckland suggests starting modestly by focusing on childhood immunizations, influenza vaccinations among persons over sixty-five, cervical smears and breast screening (Perkins et al. 2006).

Public reporting of performance measures could also motivate clinicians to turn away or deny care to potentially risky patients although there is scant rigorous evidence of this (Shekelle 2009). The most frequently cited example involves New York State, which has published hospital- and physician-level report cards on coronary artery bypass graft (CABG) surgery deaths since the early 1990s and, more recently, on coronary angioplasty outcomes. Anecdotal rumours among thoracic surgeons and interventional cardiologists, as well as limited objective evidence, suggest that public reporting has made certain New York clinicians reluctant to accept patients with relatively high mortality risks. The concern (not yet proven conclusively) is that high-risk New York residents in need of a CABG or angioplasty must seek physicians elsewhere. Ironically, CABG mortality has one of the most evidence-based, intensively validated and extensively honed risk adjustment methodologies of all performance measures (McMahon et al. 2007). If these reports of avoiding high-risk patients hold true, it would be impossible to forestall gaming behaviour among worried clinicians.

Finally, failure to risk adjust hampers attempts to engage providers in a meaningful dialogue about improving performance. Clinicians

may simply argue that unadjusted data are unfair and misrepresent their patient panels, impeding efforts to use these data to direct quality improvement activities. Distinguishing the factors that clinicians can control from those they cannot is a key aim of risk adjustment and essential to identifying productive improvement strategies.

Risk adjustment for different performance measures

The word risk is meaningless without first answering the fundamental question – risk of what? (Iezzoni 2003). In measuring health-care quality this question generates countless answers (from imminent death to satisfaction with care) across diverse health-care settings. For instance, risk adjustment for comparison of CABG death rates differs from that for consumer satisfaction with hospice care. The need for and nature of risk adjustment varies with the topic of interest.

It is necessary to acknowledge limitations in the current science of performance measurement before discussing risk adjusting performance measures. Today, numerous putative performance measures exist for diverse clinical areas and settings of care. Nonetheless, an Institute of Medicine (2006) committee review of more than 800 performance measures identified significant gaps and inadequacies in current quality measures. The scientific evidence base for specifying quality measures remains insufficient in many clinical areas. Numerous existing performance measures focus on actions or activities with limited or unproven clinical value and many concerns relate to risk adjustment and identifying at-risk patients. As Hayward (2007 p.952) observed, the field needs to: "construct performance measures that are much more nuanced and that consider patients' preferences, competing needs, and the complex circumstances of individual patients. Extensive work has shown how simplistic, all-or-nothing performance measurement can mislead providers into prioritizing low-value care and can create undue incentives for getting rid of 'bad' patients."

Growing populations of older persons with multiple co-morbid conditions are especially neglected by current disease-by-disease performance measurement approaches. Boyd et al. (2005) applied established practice guidelines (often the source of performance measures) to a hypothetical 79-year-old woman with hypertension, diabetes mellitus, osteoporosis, osteoarthritis and chronic obstructive pulmonary

disease. To meet guideline specifications, the woman would need to pursue fourteen nonpharmaceutical activities and take twelve separate medications in a regimen requiring nineteen daily drug doses. Some recommendations contradicted each other, thus endangering her overall health. There are rapidly ageing populations in many nations worldwide. Accounting for the clinical complexities of persons with multiple chronic conditions and individual preferences for care presents a major challenge for performance measurement and holds important implications for risk adjustment.

Outcome versus process measures

Performance measures often sort into two types: (i) outcomes – how patients do; and (ii) processes of care – what is done to and for patients. Outcomes generally have a clear rationale for risk adjustment. How patients do in the future is closely related to how they are doing now or did in the recent past. Risk adjustment is obviously essential for outcomes heavily influenced by patients' intrinsic clinical characteristics over which clinicians have little control. For example, gravely ill intensive treatment unit (ITU) patients are at greater risk of the outcome 'imminent death' than moderately ill patients. Researchers have developed good methods to risk adjust ITU mortality rates through years of analysing indicators of disease burden and physiological functioning (e.g. vital signs, serum chemistry findings, level of consciousness). Much of the early work on ITU risk adjustment occurred in the United States (Knaus et al. 1981, 1985 & 1991) but these models have been validated and new ones developed in nations worldwide. Methods for risk adjusting paediatric and adult ITU mortality rates are readily available e.g. the United Kingdom's Intensive Care National Audit & Research Centre (www.icnarc.org). It is critically important to validate risk adjustment methods within individual countries for the outcome 'ITU mortality'. Although basic human physiology does not vary, practice patterns (e.g. admission policies, available technologies) and patients' preferences (e.g. use of do-not-resuscitate status) certainly do. These considerations could affect associations of physiological risk factors with mortality outcomes.

Risk adjustment methods pertaining to hospitalization outcomes (primarily mortality and, increasingly, complications of care) have

been the most studied over the last thirty years. As noted above, clinically detailed risk adjustment methods for coronary artery bypass graft surgery and coronary interventions are well-developed. In the National Veterans Administration Surgical Risk Study researchers spent more than fifteen years developing risk adjustment methods using clinical variables for other selected surgical specialties (Khuri et al. 1995 & 1997). These methods are now available in the private sector through the American College of Surgeons National Surgical Quality Improvement Program (NSQIP).

This brief chapter cannot itemize the expanding number of publicly available and commercial risk adjustment methods developed to target various outcomes within differing settings of care. Suffice to say that existing risk adjustment methods differ widely in terms of their risk factor specifications, weighting schemes and validation for applications in practice settings beyond those in which they were developed (i.e. other countries with differing practice patterns), depending on the particular outcome, care environment and purpose.

It has been particularly challenging to risk adjust outcomes of routine outpatient care for performance measurement involving common chronic conditions. A number of the 146 indicators chosen for the NHS 2004 pay-for-performance initiative involved outcomes of care. Although patient attributes could certainly affect the selected outcomes, the NHS programme did not conduct formal risk adjustment. Instead, general practitioners received points for their performance between specified minimum and maximum values – an approach that Velasco-Garrido et al. (2005 p.231) describe as 'a kind of simple method for risk adjustment' (Box 3.1.3). However, this characterization is not entirely compatible with the usual goals of risk adjustment. In the example given in Box 3.1.3, if that practice's panel comprised patients with a heavy burden of co-morbid illness and difficult to control diabetes then bringing only 30% to the target blood pressure may represent a significant clinical achievement, perhaps worthy of nearly the full seventeen points. Actual risk adjustment would account for this underlying clinical complexity and pro-rate the point scheme accordingly. The NHS methods' failure to recognize these types of problems might have contributed to concerns about exception report-

Box 3.1.3 UK NHS blood pressure indicator

A maximum of seventeen points can be achieved for controlling blood pressure in diabetic patients (i.e. BP 145/85 mmHg or less). The threshold to obtain a score is 25% of patients; the maximum practically achievable has been set at 55%. A practice that achieves this target blood pressure in 55% of its diabetic patients will obtain the full score for this indicator. If the target is achieved for only 30% of the diabetic patients, the practice score for this indicator will be only 5/30, that is 2.8 points.

Source: Velasco-Garrido et al. 2005.

ing (i.e. eliminating patients from a particular quality indicator report) (Doran et al. 2006).[3]

Process measures (what is done to and for patients) can also warrant risk adjustment. Beyond patients' clinical attributes, certain process measures may require adjustment for non-clinical factors that may confound performance assessment – factors that can be 'difficult to measure and account for with risk adjustment' (Birkmeyer et al. 2006 p.189). These might include patients' psychosocial characteristics, socio-economic status and preferences for care.

Many process measures build in explicit specifications of patient characteristics that are essentially risk factors for obtaining the service. These factors act as inclusion or exclusion criteria, indicating which subset of patients qualifies to receive the process of care. For example, in the United States it is a widely accepted process mea-

[3] Family practitioners can exclude or exception-report patients for reasons including: family practitioner judges indicator inappropriate for the patient because of particular circumstances, such as terminal illness, extreme frailty or the presence of a supervening condition that makes the specified treatment of the patient's condition clinically inappropriate; patient has had an allergic or other adverse reaction to a specified medication or has another contraindication to the medication; patient does not agree to investigation or treatment (Doran et al. 2006).

sure to administer aspirin to patients admitted to hospital with acute myocardial infarction, with the stipulation that patients do not have any of a list of contraindications or exclusion criteria (Kahn et al. 2006).[4] Comparisons of the fraction of acute myocardial infarction patients receiving aspirin across hospitals must recognize that the mix of patients with contraindications may differ across facilities. Here, it is most appropriate to apply contraindication criteria individually, case-by-case (i.e. determining whether aspirin is clinically indicated for each patient). Comparisons across hospitals then focus only on those patients without contraindications and makes it unnecessary to risk adjust for conditions considered as exclusion criteria. This process appears straightforward but even panels of experts can find it challenging to specify inclusion and exclusion criteria in certain clinical contexts (Shahian et al. 2007).

Measures involving patient preferences

Process measures that require a positive action by patients (i.e. obtaining a mammogram, having a child immunized) raise special concerns. These actions are affected by education, motivation, wherewithal (e.g. financial resources, transportation, child care, time off work), preferences for care and outcomes, cultural concerns and various other factors – largely outside clinician control. Different clinicians and providers of care see different mixes of patients along these critical dimensions, raising the need for risk adjustment. For certain purposes, risk stratification might offer a more informative way to present these comparisons (see below).

The underlying goals of process-driven quality measurement initiatives carry implications for risk adjusting the performance measures. For example, health-care administrators may decree that virtually all older women should undergo mammography, regardless of their sociodemographic characteristics. Providers caring for large fractions of

[4] The Joint Commission in the United States specifies hospital performance measures widely used in federal reporting initiatives. Exclusions listed for the aspirin on admission measure are: active bleeding on arrival to the hospital or within twenty-four hours of arrival; aspirin allergy; pre-arrival use of warfarin; or other reasons documented by specified clinicians for not administering aspirin before or after admission (Kahn et al. 2006).

women who, for whatever reason (e.g. education, culture, resources), are less apt to obtain a mammogram should nonetheless be held to the same standard as other providers. In this circumstance, risk adjustment becomes moot. This stance might have merit (e.g. equity across patient subgroups) but has practical consequences. Providers that must spend resources boosting their mammography rates may neglect other issues. This also disregards the role of patient preferences, one factor considered in NHS exception reporting (Doran et al. 2006).

Patient preferences are not only an issue for process measures but also might affect some outcomes directly. Mortality rates are a prime example. According to Holloway and Quill (2007 p.802), 'mortality has been criticized as a measure of quality for years and debates about methods of risk adjustment are almost clichéd', but these debates neglect concerns about 'preference-sensitive care.' Hospitals vary widely in the use of early do-not-resuscitate orders and hospital mortality measures erroneously treat all deaths as medical failures. In 2007, Medicare launched Hospital Compare (www.hospitalcompare.hhs.gov), a web site that posts various performance measures including risk adjusted mortality rates for acute myocardial infarction and congestive heart failure for hospitals nationwide. Hospital Compare identified a hospital in Buffalo, New York, as one of the thirty-five worst American hospitals because its mortality rate for congestive heart failure between July 2005 and June 2006 was 4.9% more than the national mean. The hospital reviewed medical records of these deaths and found that eleven decedents (about 40% of the total) were in hospice or receiving only palliative care treatment at patients' requests (Holloway & Quill 2007). More than twenty years after its initial problematic data release (Box 3.1.2), Medicare's risk adjustment method still did not account for patients' preferences for end-of-life care. Some initiatives that report hospital mortality rates exclude all hospice patients from these calculations. This eliminates the need to risk adjust for this patient preference, assuming that all patients with early do-not-resuscitate orders are in hospice (which may not always occur) (Holloway & Quill 2007). In the United Kingdom, whether patients were admitted to palliative care units was recently added to the list of risk factors for computing hospital standardized mortality ratios (Dr Foster Intelligence 2007).

Composite measures

As detailed in Chapter 3.4, there is increasing interest in combining diverse individual performance measures to produce composites or summary assessments of quality-related performance. A conceptual justification for this approach is the complexity of quality, comprised of multifaceted dimensions. A practical impetus for producing composite measures involves common statistical realities – small sample sizes of patients for clinicians, hospitals or other units of interest; and the relative rarity of many targeted single events, such as deaths. The simplicity offered by a single number or score has led some groups to propose the creation of composite performance measures that cut across Donabedian's (1980) classic triad of quality measurement dimensions: outcomes, processes and structures of care (Shahian et al. 2007).

Despite the appeal of simple summary scores the production of composite ratings raises important methodological questions, including whether individual measures within the composite require risk adjustment. The construction of composite measures is complicated and stokes fears about complex statistical arguments masking opportunities for manipulation or misinterpretation. Since September 2001, the NHS in England has published annual star ratings for acute care hospitals, using composite scores to assign hospitals to one of four levels: from zero to three stars. Jacobs et al. (2006) used data from these star ratings to explore the stability of these composite hospital rankings across different methodological choices. They found considerable instability in hospitals' positions in league tables. Beyond those overarching problems, details of individual measures can be lost within the composite. For instance, coronary artery bypass graft mortality is one of the many indicators combined in the star rating composite but these mortality rates are not risk adjusted. Producers of the star ratings aim to ease concerns about these unadjusted mortality figures by comparing institutions within different classes of hospitals – ostensibly a broad attempt to control for patients' risks.

The Society of Thoracic Surgeons (STS) quality measurement task force in the United States has demonstrated the complexities of producing composite measures while paying detailed attention to risk adjustment (O'Brien et al 2007; Shahian et al. 2007). Table 3.1.1 shows

Table 3.1.1 ***Individual measures and domains in the STS composite quality score***

Operative care domain
• use of at least one internal mammary artery graft
Perioperative medical care domain
• preoperative beta blockers
• discharge beta blockers
• discharge antiplatelet medication
• discharge antilipid medication
Risk adjusted mortality domain
• operative mortality
Risk adjusted major morbidity domain
• prolonged ventilator (> 24 hours)
• deep sternal wound infection
• permanent stroke
• renal insufficiency
• reoperation

Source: O'Brien et al. 2007

the eleven performance measures selected for producing the composite score. Analysts defined and estimated six different risk adjusted measures to add to their summary model, one for each of the six items requiring risk adjustment. Using clinical data from a large STS data set representing 530 providers, their multivariate random-effects models estimated true provider-specific usage rates for each process measure and true risk-standardized event rates for each outcome. Further analyses suggested that each of the eleven items provided complementary rather than redundant information about performance (O'Brien et al. 2007). Despite their extensive analyses, the STS investigators acknowledge the need to monitor the stability of their composite scores over time; and sensitivity to various threats, such as nonrandomly missing data used for risk adjustment. Future research must explore not only the benefits and drawbacks of composite performance measures but also the role that risk adjustment of individual indicators plays in summary rankings.

Conceptualizing risk factors

The development and validation of credible risk adjustment methods requires substantial time and resources. This chapter does not have the space to describe the steps needed to complete this process but looks briefly at three major issues pertaining to the development of risk adjustment methods: (i) choice of risk factors; (ii) selection and implications of data sources; and (iii) overview of statistical methods.

The essential first step in risk adjusting performance measures involves a thorough understanding of the measure and its validity as a quality indicator. The next step is to develop a conceptual model identifying patient factors that could potentially affect the targeted outcome or process of care. Table 3.1.2 suggests various patient-risk factors grouped along different dimensions although additional attributes could apply to the wide range of potential performance measurement topics and settings of care (Iezzoni 2003).

Initially analysts should develop this conceptual model independently of practical concerns, particularly about the availability of data. Pertinent characteristics and their relative importance as risk factors vary across different performance measures. For example, indicators of acute physiological stability (e.g. vital signs, serum electrolytes, arterial oxygenation) are critical for assessing risk of imminent ITU death but less important for evaluating consumer satisfaction with health plans. It is impossible to risk adjust for all patient dimensions. Nevertheless, it is essential to know what potentially important factors have not been included in risk adjustment. This assists in interpreting comparisons of performance measures across clinicians, hospitals or other providers – attributing residual differences in performance to their root cause (i.e. unmeasured patient characteristics versus other factors).

The selection of potential risk factors can prove controversial, especially items chosen as potential proxies when data about a particular risk factor are unavailable. For example, in England Dr Foster Intelligence produces an annual guide that ranks acute hospital trusts by standardized mortality ratios. Recently, analysts added to their risk adjustment model – each patient's previous emergency admissions within the last twelve months. Presumably, this aims to capture something about the patients' clinical stability and status of chronic illnesses. However, this risk factor could be confounded with the very

Table 3.1.2 ***Potential patient risk factors***

Demographic characteristics
- age
- sex/gender
- race and ethnicity

Clinical factors
- acute physiological stability
- principal diagnosis
- severity of principal diagnosis
- extent and severity of co-morbidities
- physical functioning
- vision, hearing, speech functioning
- cognitive functioning
- mental illness, emotional health

Socio-economic/psychosocial factors
- educational attainment, health literacy
- language(s)
- economic resources
- employment and occupation
- familial characteristics and household composition
- housing and neighbourhood characteristics
- health insurance coverage
- cultural beliefs and behaviours
- religious beliefs and behaviours, spirituality

Health-related behaviours and activities
- tobacco use
- alcohol, illicit drug use
- sexual practices ('safe sex')
- diet and nutrition
- physical activity, exercise
- obesity and overweight

Attitudes and perceptions
- overall health status and quality of life
- preferences, values and expectations for health-care services

quantity that standardized mortality ratios aim to highlight – quality of care. Patients may have more emergency readmissions because of poor quality of care (e.g. premature discharges, inadequate care) during prior admissions at that same hospital. In this instance, control-

ling for frequent readmissions might give hospitals credit for sicker patients rather than highlighting the real problem. Documentation from Dr Foster Intelligence indicates that 'adjustments are made for the factors that are found by statistical analysis to be significantly associated with hospital death rates' (Dr Foster Intelligence 2007). However, as this example suggests, choosing risk factors based only on statistical significance could mask mortality differences related to poor hospital care. Risk factors – and their precise specification (e.g. if using a proxy) – should have clear conceptual justification relating to elucidating provider quality.

Some risk adjustment methods employ processes of care as risk factors, generally as proxies for the presence or severity of disease. Examples include use of certain pharmaceuticals or procedures generally reserved for very ill patients (e.g. tracheostomy, surgical insertion of gastric feeding tube). These processes might have clinical validity as indicators of patients' future risks but in the context of performance measurement for pay for performance or public reporting they are potentially susceptible to manipulation or gaming (see below). These concerns argue against the use of processes of care as risk factors.

Data options and implications

Inadequate information is the biggest practical impediment to risk adjustment. Required information may be simply unavailable or too costly or infeasible to obtain. The conceptual ideal is to have complete information on all potential risk factors (Table 3.1.2) but that goal is not readily unattainable. Therefore, risk adjustment today is inevitably an exercise in compromise, with important implications for interpreting the results. The three primary sources of data for risk adjustment, each with advantages and disadvantages, are now described in more detail.

Administrative data

Administrative data are the first primary source. By definition, they are generated to meet some administrative purpose such as claims submitted for billing or records required for documenting services. The prototypical administrative data record contains a patient's administrative identification and demographic information; one or more diagnoses coded using some version or variant of the WHO

ICD; procedures coded using some local coding classification (unlike ICD, which is used in some form worldwide, there is no universal coding system); dates of various services; provider identifiers; and perhaps some indication of costs or charges, depending on the country and setting of care. To maximize administrative efficiency these records are ideally computerized, submitted electronically and relatively easy to obtain and analyse.

Administrative data offer the significant advantage of ready availability, ease of access and relatively low acquisition costs. Required data elements are typically clearly defined and theoretically recorded using consistent rules, ostensibly making the data content comparable across providers. Administrative records also typically cover large populations, such as all persons covered by a given health plan or living in a specific geographical area. Uniform patient identifiers enable analysts to link records relating to individual patients longitudinally over time (e.g. as when creating the Dr Foster variable relating to prior emergency admissions). Some countries (e.g. United States, United Kingdom, Sweden) have spent considerable resources upgrading their electronic administrative data reporting in anticipation of using this information to manage their health-care systems more effectively (Foundation for Information Policy Research 2005).

However, significant disadvantages can make risk adjustment methods derived from administrative data immediately suspect. Payment-related incentives can skew data content especially when providers produce administrative records to obtain reimbursement. The most prominent example in the United States involved inaccurate reporting of diagnosis codes when Medicare first adopted DRG-based prospective payment. Coding audits found that hospitals engaged in 'DRG creep' (Hsia et al. 1988; Simborg 1981) by assigning diagnoses not supported by medical record evidence but likely intended to move patients into higher-paying DRGs. Inconsistencies and inaccuracies in the assignment of ICD codes across providers can compromise comparisons of their performance using administrative data. Systematic biases across hospitals in under- or over-reporting diagnoses could compromise comparisons. For example, in England, foundation acute hospital trusts have lower rates of uncoded data than other acute trusts. They have prioritized the improvement of coding accuracy and timeliness by investing in training; hiring additional data coders and health information managers; and encouraging coding directly from

medical records rather than discharge summaries (Audit Commission 2005). In 2004/2005, the average acute hospital admission in England received only 2.48 coded diagnoses, compared with just over three diagnoses in Australia and six in the United States (Audit Commission 2005 p.47).

Hospital coding of diagnoses raises additional questions about comparing quality performance. Romano et al. (2002) examined results from a reabstraction of 991 discectomy cases admitted to California hospitals. The original hospital codes displayed only 35% sensitivity for identifying any complication of care found during reabstraction (i.e. the gold standard). Under-reporting was markedly worse at hospitals calculated to have lower risk adjusted complication rates. Undercoding extended beyond serious complications to more mild conditions, such as atelectasis, post-haemorrhagic anaemia and hypotension. One study from Canada examined the concordance between medical records and administrative data for conditions included in the Charlson co-morbidity index commonly used in risk adjustment (e.g. Dr Foster uses Charlson co-morbidities in its standardized hospital mortality ratios). Administrative data under-reported ten co-morbidities but slightly over-reported diabetes, mild liver disease and rheumatological conditions (Quan et al. 2002 pp. 675-685).

There are also reservations about the clinical content of ICD codes. Although these aim to classify the full range of diseases and various health conditions that affect humans, they do not capture the critical clinical parameters associated with illness severity (e.g. arterial oxygenation level, haematocrit value, extent and pattern of coronary artery occlusion); nor do they provide insight into functional impairments and disability (see WHO 2001 for that purpose).[5] In the United States[6] these reservations have prompted more than a decade of research controversy as Medicare has tried to produce clinically cred-

[5] Representatives from numerous nations participated in specification of WHO's ICF (revision of the International Classification of Impairments, Disabilities and Handicaps). Nonetheless, it is unclear how systematically this is used in administrative data reporting around the world. It does not appear on administrative records required by Medicare or major health insurers in the United States.

[6] United States has switched to ICD-10 for reporting causes of death but still uses a version of ICD-9 specifically designed by American clinicians for morbidity reporting – ICD-9-CM (http://www.eicd.com/EICDMain.htm).

ible risk adjusted mortality figures without the considerable expense of widespread data gathering from medical records. For the Hospital Compare web site, Medicare contracted with researchers at Yale University to develop administrative data-based risk adjustment algorithms for acute myocardial infarction and congestive heart failure mortality within thirty days of hospital admission and to validate the results against methods using detailed clinical information abstracted from medical records (Krumholz et al. 2006). The correlation of standardized hospital mortality rates calculated with administrative versus clinical data was 0.90 for acute myocardial infarction and 0.95 for congestive heart failure. These findings and the results of other statistical testing suggested that the administrative data-based models were sufficiently robust for public reporting.

Cardiac surgeons remain sceptical about whether administrative data can produce meaningful risk adjustment for coronary artery bypass graft hospital mortality rankings. Shahian et al. (2007a) examined this question using detailed clinical data gathered during coronary artery bypass graft admissions in Massachusetts hospitals. The administrative mortality model used risk adjustment methods promulgated by the federal AHRQ and built around all patient refined DRGs (APR-DRGs).[7] The researchers also tested differences between examining in-hospital versus thirty-day post-admission mortality and the implications of using different statistical methodologies (i.e. hierarchical versus standard logistic regression models). At the outset, one major problem was cases misclassified as having had isolated coronary artery bypass graft surgery – about 10% of the administratively identified coronary artery bypass graft cases had some other simultaneous but poorly specified surgery (another subset had concomitant valve surgery). Risk adjusted outcomes varied across the two data sources because of both missing risk factors in the administrative models and case misclassification.

Shahian et al's study (2007a) also highlighted difficulties determining the timing of in-hospital clinical events using coded data. This raises its own set of problems. Administrative hospital discharge data

[7] All APR-DRGs were developed by 3M Health Information Systems (Wallingford, CT, USA) to predict two different outcomes: resource use during hospital admissions and in-hospital mortality. The two models use different weighting schemes for the predictor variables (primarily ICD-9-CM discharge diagnoses) and produce different scoring results.

generally have not differentiated diagnoses representing post-admission complications from clinical conditions existing on admission. A tautology could occur if administrative data based risk adjusters use codes indicating virtual death (e.g. cardiac arrest) to predict death, raising the appearance that the model performed well statistically (e.g. producing artifactually high R-squared values or c statistics). Lawthers et al. (2000) looked at the timing of secondary hospital discharge diagnoses by reabstracting over 1200 medical records from hospitalizations in California and Connecticut. Among surgical cases they found many serious secondary diagnosis codes representing conditions that occurred following admission, including 78% of deep vein thrombosis or pulmonary embolism diagnoses and 71% of instances of shock or cardiorespiratory arrest. In our work, discharge abstract-based risk adjusters were generally equal or better statistical predictors of in-hospital mortality than measures derived from admission clinical findings (Iezzoni 1997). Not surprisingly, the administrative risk adjustment models appeared over-specified in the coronary artery bypass graft study (Shahian et al. 2007a).[8] However, even more important than this statistical concern is the possibility that risk adjusters that give credit for potentially lethal in-hospital events might mask the very quantity of ultimate interest – quality of care.

Since 1 October 2008, Medicare has required hospitals in the United States to indicate whether each coded hospital discharge diagnosis was present on admission (POA) or occurred subsequently (e.g. in-hospital complication) for hospitalized beneficiaries. A POA indicator would allow risk adjustment methods to use only those conditions that patients brought with them into the hospital, potentially isolating diagnoses caused by substandard care (Zhan et al. 2007). POA flags could substantially increase the value of hospital discharge diagnosis codes for measuring quality performance. However, California and New York implemented POA flags for discharge diagnoses years ago and subsequent studies have raised questions about the accuracy of these indicators (Coffey et al. 2006).

[8] Over-specification could occur when post-operative events virtually synonymous with death (e.g. cardiac arrest) are used in the risk adjustment models. Models containing such rare but highly predictive events may not validate well (e.g. when applied to other data sets or a portion of the model development data set withheld for validation purposes), thus indicating model over-specification.

Medical records or clinical data

The second primary source of risk factor information is medical records or electronic systems containing detailed clinical information in digital formats (e.g. electronic data repositories). The primary benefit of these data is clinical credibility. This clinical face validity is essential for the acceptance of risk adjustment methods in certain contexts, such as predicting coronary artery bypass graft mortality (Shahian et al. 2007a) and deaths following other operations (Khuri et al. 1995 &1997). In certain instances (e.g. when risk adjusting nursing home or home health-care outcomes) coded administrative data provide insufficient clinical content and validity. ICD diagnosis codes do not credibly capture clinical risk factors in these non-acute care settings where patients' functional status typically drives outcomes.

Abstracting information from medical records is expensive and raises other important questions. To ensure good data quality and comparability, explicit definitions of the clinical variables and detailed abstraction guidelines are required when collecting clinical information across providers. Gathering extensive clinical information for performance measurement may demand extensive training and monitoring of skilled staff to maintain data quality. It is hoped that electronic medical records, automated databases and electronic data repositories will eventually ease these feasibility concerns. For instance, Escobar et al. (2008) linked patient-level information from administrative data sources with automated inpatient, outpatient and laboratory databases to produce risk adjusted inpatient and thirty-day post-admission mortality models. In order to avoid confounding risk factors with possible quality shortfalls, they included only those laboratory values obtained within the twenty-four hours preceding hospitalization in their acute physiology measure. It is beyond the scope of this chapter to describe global efforts to develop electronic health information systems but countries worldwide are investing heavily in creating electronic health information infrastructures that are interoperable (i.e. allowing data-sharing readily across borders and settings of care) (Kalra 2006 & 2006a). It may even become possible to download detailed clinical data directly from these electronic systems to support risk adjustment.

Electronic records have obvious advantages (chiefly legibility) but their medical record content may not advance far beyond that of paper

records without significant changes in the documentation practices of clinicians. Especially in outpatient settings, medical records have highly variable completeness and accuracy; lengthy medical records in academic medical centres may contain notations from multiple layers of clinicians, sometimes containing contradictory information (Iezzoni et al. 1992). This may partly explain why it is more challenging to capture some variables more reliably than others. For instance, reabstractions of clinical data from the Veterans Affairs National Surgical Quality Improvement Project in the United States found 97.4% exact agreement for abstracting the anaesthetic technique used during surgery; 94.9% for whether the patient had diabetes and 83.4% for whether the patient experienced dyspnea (Davis et al. 2007). Electronic medical records may contain templates with explicit slots for documenting certain data elements; some may even provide completed templates (e.g. clinical information about presumed findings from physical examinations) that allow clinicians to modify automated data entries to reflect individual clinical circumstances. Not surprisingly, concerns arise about the accuracy of such automated records. In the United States, anecdotal reports question whether clinicians actually perform complete physical examinations or just accept template data without validating the information.

Something akin to code creep might also arise when risk adjustment uses detailed clinical information as even these risk adjusters are susceptible to potential manipulation. For example, anecdotal observations suggested that routine blood testing of patients increased after a severity measure (based on extensive medical record reviews and numerous clinical findings) was mandated for publicly reporting mortality and morbidity rates at Pennsylvania hospitals in 1986. Observers have argued about whether reporting of significant clinical risk factors increased in New York following the public release of surgeon-specific coronary artery bypass graft mortality rates. Some manipulation is impossible to detect using routine auditing methods (e.g. re-review of medical records). For example, one risk factor in New York's coronary artery bypass graft mortality model is patients' physical functional limitations caused by their cardiovascular disease. Physicians make this assessment in their offices or at the bedside by questioning and examining patients. Physicians may document functional impairments in the medical record in order to exaggerate their patients' true deficits

and make them appear sicker. The only way to detect this problem is by independently re-examining patients – a costly and infeasible undertaking.

Information in administrative and medical records is always susceptible to manipulation but audits to monitor and ensure data integrity and quality are costly and sometimes impossible. The degree of motivation for gaming data reporting relates directly to clinicians' perceptions of whether risk adjusted performance measures are used punitively or unfairly. Once data are systematically and significantly gamed, they generally lose their utility for risk adjustment.

Information directly from patients or consumers

The third, and a popular, source of information is patients themselves, especially when performance measures target patients' perceptions (e.g. satisfaction with care, self-reported functional status). Patients are the only valid source of information about their views of their health-care experiences. Extensive research suggests that persons who say they are in poorer health systematically report lower levels of satisfaction with their health care than healthier individuals. Therefore, surveys asking about satisfaction typically contain questions about respondents' overall health which are then used to risk adjust the satisfaction ratings. Patients do not generally have strong motivations for gaming or manipulating their responses although studies suggest that many patients are reluctant to criticize their clinical caregivers.

Gathering data directly from patients has downsides beyond the considerable expense and feasibility challenges. Patients are not completely reliable sources of information about their specific health conditions or health service use – faulty memories, misunderstanding and misinformation compromise accuracy. Language problems, illiteracy, cultural concerns, cognitive impairments and other psychosocial issues complicate efforts to obtain information directly from patients. Education, income level, family supports, housing arrangements, substance abuse, mental illness and other such factors can affect certain outcomes of care but questions about these generate extreme sensitivities. Concerns about the confidentiality of data and sensitivity of certain issues make it infeasible to gather information on some important risk factors.

Response rates are critical to the validity of results and certain subpopulations are less likely to complete surveys.[9] Unless surveys are administered in accessible formats, persons with certain types of disabilities might be unable to respond. Furthermore, anecdotal reports from some American health insurers suggest that their enrollees are growing impatient with being surveyed about their health-care experiences. Even insurers with affluent enrollees (a population relatively likely to complete surveys) report that many of their subscribers no longer respond. The relatively few completed surveys that are available thus provide information of a highly suspect quality due to possible respondent bias.

Statistical considerations

Researchers developed the earliest generation of severity measures around thirty years ago, before large data sets containing information across numerous providers became available. After identifying risk factors, clinical experts used their judgment and expertise to specify weights (i.e. numbers indicating the relative importance of different risk factors for predicting the outcome of interest) that would be added or manipulated in some other way to produce risk scores. Now that large databases contain information from many providers, researchers can apply increasingly sophisticated statistical modelling techniques to produce weighting schemes and other algorithms to calculate patients' risks. Other chapters provide details about specific statistical methods (e.g. hierarchical modelling, smoothing techniques that attempt to improve predictive performance and recognize various sources of possible variation) but several points are emphasized here.

First, optimal risk adjustment models result from an iterative combination of clinical judgment and statistical modelling. Clinicians specify variables of interest and hypothesized relationships with the dependent variable (e.g. positive or negative correlations) and methodologists confirm whether the associations are statistically significant and satisfy hypotheses. Final models should retain only clinically credible factors that are not confounded with the ultimate goal of perfor-

[9] Surveys of Medicare beneficiaries' perceptions of health-care experiences suggest that certain subpopulations are especially unlikely to respond, e.g. older individuals; people with disabilities; women; racial and ethnic minorities; those living in geographical areas with relatively high rates of poverty and low education.

mance measurement – assessing quality of care. Thus, the creation of a risk adjustment method is a multidisciplinary effort. At a minimum this involves clinicians interacting with statisticians but may require experts in information systems and data production (e.g. medical record and coding personnel); quality improvement; survey design; and management. Analysts should avoid the urge to data dredge. With large databases and fast powerful computers, it is tempting to let the computer specify the risk adjustment algorithm (e.g. select variables) with minimal human input. Users of risk adjustment models should remain sceptical until models are confirmed as clinically credible and statistically validated, preferably on a data set distinct from that used to derive the model.

Second, models developed in one country may not necessarily transfer easily to another. Differences in practice patterns, patient preferences, data specifications and other factors could compromise validity and statistical performance in different settings. Clinicians and methodologists should examine both clinical validity and statistical performance before using models developed elsewhere.

Third, summary statistical performance measures (e.g. R-squared and c statistics) suggest how well risk adjustment models perform at predicting the outcomes of interest or discriminating between patients with and without the outcome. These measures are attractive because they summarize complex statistical relationships in a single number. However, it can be misleading to look only at (for example) relative R-squared values to choose a risk adjustment model. Quirks of the database or selected variables can inflate summary statistical performance measures and experienced analysts know that some data sets are easier to manipulate (e.g. because of the range or distribution of values of variables). Sometimes available predictor (independent) variables may be confounded with the outcome (dependent) variable. An example of this was noted above: when predicting hospital mortality, diagnosis codes that indicate conditions that occurred following admission can elevate c-statistics but obviously confound efforts to find quality problems. Summary statistical performance measures do not indicate how well risk adjustment models predict outcomes for different subgroups of patients. Therefore, decision-makers choosing among risk adjustment methods ideally should not simply search for the highest R-squared or c statistic but should also consider clinical validity and ability to isolate quality deficits.

Finally, other policy considerations may affect decisions about how to risk adjust comparisons of performance measures across practitioners, institutions or other units of interest. Statistical techniques control for the effects of risk factors and allow analysts to ignore these patient characteristics as the explanation for observed outcome differences. However, situations can arise where policy-makers suspect that quality also varies by critical patient characteristics, such as race or social class. Risk stratification can prove useful if the mix of these characteristics differs across the groups being compared (e.g. clinician practices, hospitals) as it examines the performance within strata (i.e. groups) of patients defined by the specific characteristic. Such analyses are especially important when the specific patient attribute has important social policy implications, such as ensuring equitable care across subpopulations.

An example from the United States highlights how risk stratification might work. Research indicates that African-American women are less likely than white women to obtain mammograms. Multiple factors likely contribute to this disparity, including differentials in educational level, awareness of personal breast cancer risks and women's preferences. If two health plans have different proportions of black and white enrollees then risk adjustment controlling for race will not reveal whether the health plans have similar or divergent mammography rates for black and white women. It might also mask a plan's especially poor mammography performance among its black enrollees. In this instance, analysts should perform race-stratified comparisons – looking at mammography rates for black women and for white women respectively across the two plans.

When is risk stratification indicated? The answer underscores the critical importance of understanding the context in which the risk adjusted information will be used and having a conceptual model of the relationships between a given performance measure and various potential risk factors. Risk stratification is desirable when analysts believe that a policy-sensitive patient characteristic (e.g. race, social class) is an important risk factor but could also reflect differences in the treatments patients receive (i.e. quality of care). In this situation, analyses that begin with risk stratification can provide valuable insight. If performance is similar for different comparison groups (e.g. health plans, hospitals) within each patient stratum, then analysts could reasonably combine patients across strata and risk adjust for that char-

acteristic, assuming that the conceptual model provides a valid causal rationale for including that characteristic among the risk factors.[10]

Plea for transparency

As suggested above, risk adjustment is a complicated business – literally so in some health-care marketplaces such as the United States. Many proprietary organizations, health information vendors and others promote or sell their own risk adjustment methodologies for a range of purposes. Policy-makers should be sceptical of marketing claims and would be wise to request details and rigorously evaluate methods to examine whether: they are clinically sound; important risk factors are missing; the data used are sufficiently sound; and the statistical methods are reasonable. However, it is often difficult (if not impossible) to gain access to important details about proprietary methods

When performance measures are either legally mandated or de facto required, policy-makers should consider stipulating that vendors make complete details of the risk adjustment method available for external scrutiny. An ideal strategy would place these methods in the public domain and ensure that they meet minimal explicit standards of clinical credibility and statistical rigour. An external, independent and objective body could operate an accreditation process through a standard battery of evaluations to establish whether the methods meet established explicit criteria of clinical validity and methodological soundness. Analysts should compare competing risk adjustment methods by applying them to the same database as results obtained from different data sets are not truly comparable. Testing would identify not only what the methods adjust for but also what they exclude. Information on critical missing risk characteristics could appear alongside comparisons of risk adjusted performance measures to highlight factors (other than quality) that might explain differences across the units being compared.

[10] In the United States, many analysts routinely include race and ethnicity among the predictor variables in modelling a wide range of outcomes (dependent variables). Scientific evidence rarely makes direct causal links between race and ethnicity and outcomes used in performance measurement, other than as perhaps a proxy for social disadvantage (e.g. poor education, low income) or disparate quality of care. Obviously, this raises serious questions about automatic inclusion of race and ethnicity in risk adjustment models for performance measures.

Commercial vendors of risk adjustment methods will argue that putting their products into the public domain will destroy their ability to market their product and fund future developments. This contention has merit and carefully designed policies must balance private sector interests with public needs. However, a method that is mandated for widespread use should be transparent – especially if the results will be publicized. Information produced via opaque methods could compromise the goal of motivating introspection, change and quality improvement.

Conclusions

Risk adjustment is an essential tool in performance measurement. Many risk adjustment methods are now available for users to apply to their own health-care settings, after preliminary testing. However, differences in practice patterns and other factors mean that methods developed in one environment may not transfer directly to other health-care delivery systems. Methods created in resource intensive settings (e.g. the United States) may not readily apply to less technologically driven systems but it may be possible to recalibrate or revise existing risk adjusters to suit local health-care environments. This will be less costly that developing entirely new risk adjustment methods.

Inadequate data sources pose the greatest challenge to risk adjustment. No data source can ever contain information on every personal and clinical attribute that could affect health-care outcomes and unmeasured patient characteristics will always contribute to differences in patient outcomes. Improving clinical data systems – and their linkage with large, population-based administrative records – offers the greatest potential for advancing risk adjustment.

These realities should not deter policy-makers but simply heighten caution about interpreting and using the results, for example when employing risk adjusted performance measures in pay-for-performance programmes or public quality reporting initiatives. Performance measures that are labelled 'risk adjusted' (even with inadequate methods) can engender a false sense of security about the validity of results. Depending on the nature of unmeasured risk factors, it may not be realistic or credible to hold clinicians or other providers fully accountable for performance differences.

Despite these complexities, there are substantial problems associated with *not* risk adjusting. Consumers could receive misleading information; providers might strive to avoid patients perceived as high risk; and any productive dialogue about improving performance could be compromised. Nonetheless, science cannot guarantee perfect risk adjustment and therefore decisions about applying these methods will engender controversy. It is likely that legitimate arguments for and against the use of methods with inevitable shortcomings will continue and policy-makers will need to weigh up the competing arguments when deciding on the appropriate use of risk-adjusted data.

References

Anonymous (1864). 'Untitled. Response to letter by William Farr.' *Medical Times and Gazette:* 13 February 1864, pp. 187–188.

Audit Commission (2005). *Early lessons from payment by results*. London: Audit Commission.

Birkmeyer, JD. Kerr, EA. Dimick, JB (2006). Improving the quality of quality measurement. In: *Performance measurement. Accelerating improvement*. Institute of Medicine Committee on Redesigning Health Insurance Performance Measures, Payment, and Performance Improvement Programs, Washington, DC: The National Academies Press.

Boyd, CM. Darer, J. Boult, C. Fried, LP. Boult, L. Wu, AW (2005). 'Clinical practice guidelines and quality of care for older patients with multiple co-morbid diseases: implications for pay for performance.' *Journal of the American Medical Association,* 294(6): 716–724.

Brinkley, J (1986). 'US releasing lists of hospitals with abnormal mortality.' *New York Times,* 12 March 1986, Sect. A.

Bristowe, JS (1864). 'Hospital mortality.' *Medical Times and Gazette,* 30 April 1864, pp. 491–492.

Bristowe, JS. Holmes, T (1864). *Report on the hospitals of the United Kingdom. Sixth report of the medical officer of the Privy Counci, 1863*. London, UK: George E. Eyre and William Spottiswoode for Her Majesty's Stationery Office.

Casalino, LP. Alexander, GC. Jin, L. Konetzka, RT (2007). 'General internists' views on pay-for-performance and public reporting of quality scores: a national survey.' *Health Affairs,* 26(2): 492–499.

Coffey, R. Milenkovic, M. Andrews, RM (2006). *The case for the present-on-admission (POA) indicator*. Washington, DC (Agency for Healthcare Research and Quality HCUP Methods Series Report).

Davis, CL. Pierce, JR. Henderson, W. Spencer, C. Tyler, DC. Langberg, R. Swafford, J. Felan, GS. Kearns, MA. Booker, B (2007). 'Assessment of the reliability of data collected for the Department of Veterans Affairs National Surgical Quality Improvement Program.' *Journal of the American College of Surgeons,* 204(4): 550–560.

Donabedian, A (1980). *Explorations in quality assessment and monitoring*. Ann Arbor, MI: Health Administration Press.

Doran, T. Fullwood, C. Gravelle, H. Reeves, D. Kontopantelis, E. Hiroeh, U. Roland, M (2006). 'Pay-for-performance programs in family practices in the United Kingdom.' *New England Journal of Medicine,* 355(4): 375–384.

Dr Foster Intelligence (2007) [website]. *Dr Foster hospital guide: methodology for key analyses*. London: Dr Foster Intelligence (http://www.drfoster.co.uk/hospitalGuide/methodology.pdf).

Escobar, GJ. Greene, JD. Scheirer, P. Gardner, MN. Draper, D. Kipnis, P (2008). 'Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases.' *Medical Care,* 46(3): 232–239.

Forrest, CB. Villagra, VV. Pope, JE (2006). 'Managing the metric vs managing the patient: the physician's view of pay for performance.' *American Journal of Managed Care*, 12(2): 83–85.

Foundation for Information Policy Research (2005). *Healthcare IT in Europe and North America*. Sandy: National Audit Office.

Hayward, RA (2007). 'Performance measurement in search of a path.' *New England Journal of Medicine,* 356(9): 951–953.

Holloway, RG. Quill TE (2007). 'Mortality as a measure of quality: implications for palliative and end-of-life care.' *Journal of the American Medical Association*, 298(7): 802–804.

Hsia, DC. Krushat, WM. Fagan, AB. Tebbutt, JA. Kusserow, RP (1988). 'Accuracy of diagnostic coding for Medicare patients under the prospective-payment system.' *New England Journal of Medicine*, 318(6): 352–355.

Iezzoni, LI (1997). 'The risks of risk adjustment.' *Journal of the American Medical Association,* 278(19): 1600–1607.

Iezzoni, LI (2003). *Risk adjustment for measuring health care outcomes. Third edition.* Chicago, IL: Health Administration Press.

Iezzoni, LI. Restuccia, JD. Shwartz, M. Schaumburg, D. Coffman, GA. Kreger, BE. Butterly, JR. Selker, HP (1992). 'The utility of severity of illness information in assessing the quality of hospital care. The role of the clinical trajectory.' *Medical Care,* 30(5): 428–444.

Institute of Medicine Committee on Redesigning Health Insurance Performance Measures, Payment, and Performance Improvement Programs (2006).

Performance measurement. Accelerating improvement. Pathways to quality health care. Washington, DC: National Academies Press.

Jacobs, R. Goddard, M. Smith, PC (2006). *Public services: are composite measures a robust reflection of performance in the public sector?* York, UK: Centre for Health Economics (CHE Research Paper 16).

Kahn, CN 3rd. Ault, T. Isenstein, H. Potetz, L. Van Gelder, S (2006). 'Snapshot of hospital quality reporting and pay-for-performance under Medicare.' *Health Affairs,* 25(1): 148–162.

Kalra, D (2006). *eHealth Consortium 2007. Memorandum of understanding.* (http://www.ehealthinitiative.eu/pdf/Memorandum_of_Understanding.pdf).

Kalra, D (2006a). 'Electronic health record standards.' *Methods of Information in Medicine,* 45(Suppl. 1): 136–144.

Khuri, SF. Daley, J. Henderson, W. Barbour, G. Lowry, P. Irvin, G. Gibbs, J. Grover, F. Hammermeister, K. Stremple, JF (1995). 'The National Veterans Administration Surgical Risk Study: risk adjustment for the comparative assessment of the quality of surgical care.' *Journal of the American College of Surgeons,* 180(5): 519–531.

Khuri, SF. Daley, J. Henderson, W. Hur, K. Gibbs, JO. Barbour, G. Demakis, J. Irvin, G 3rd. Stremple, JF. Grover, F. McDonald, G. Passaro, E Jr. Fabria, PJ. Spencer, J. Hammermeister, K. Aust, JB (1997). 'Risk adjustment of the postoperative mortality rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs Surgical Risk Study.' *Journal of the American College of Surgeons,* 185(4): 315–327.

Knaus, WA. Draper, EA. Wagner, DP. Zimmerman, JE (1985). 'APACHE II: a severity of disease classification system.' *Critical Care Medicine,* 13(10): 818–829.

Knaus, WA. Wagner, DP. Draper, EA. Zimmerman, JE. Bergner, M. Bastos, PG. Sirio, CA. Murphy, DJ. Lotring, T. Damiano, A (1991). "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults.' *Chest,* 100(6): 1619–1636.

Knaus, WA. Zimmerman, JE. Wagner, DP. Draper, EA. Lawrence, DE (1981). 'APACHE – acute physiology and chronic health evaluation: a physiologically based classification system.' *Critical Care Medicine,* 9(8): 591–597.

Krumholz, HM. Normand, SL. Galusha, DH. Mattera, JA. Rich, AS. Wang, Y (2006). *Risk-adjustment models for AMI and HF 30-day mortality – methodology*. Washington, DC: Centers for Medicare & Medicaid Services.

Langenbrunner, JC. Orosz, E. Kutzin, J. Wiley, MM (2005). Purchasing and paying providers. In: Figueras, J. Robinson, R. Jakubowski, E (eds.).

Purchasing to improve health systems performance. New York, NY: Open University Press.

Lawthers, AG. McCarthy, EP. Davis, RB. Peterson, LE. Palmer, RH. Iezzoni, LI (2000). 'Identification of in-hospital complications from claims data. Is it valid?' *Medical Care,* 38(8): 785–795.

Lilford, R. Mohammed, MA. Spiegelhalter, D. Thomson, R (2004). 'Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma.' *Lancet,* 363(9415): 1147–1154.

McMahon, LF Jr. Hofer, TP. Hayward RA (2007). 'Physician-level P4P – DOA? Can quality-based payment be resuscitated?' *American Journal of Managed Care,* 13(5): 233–236.

Nightingale, F (1863). *Notes on hospitals. Third edition*. London: Longman, Green, Longman, Roberts and Green.

O'Brien, SM. Shahian, DM. DeLong, ER. Normand, SL. Edwards, FH. Ferraris, VA. Haan, CK. Rich, JB. Shewan, CM. Dokholyan, RS. Anderson, RP. Peterson, ED (2007). 'Quality measurement in adult cardiac surgery: part 2 – Statistical considerations in composite measure scoring and provider rating.' *Annals of Thoracic Surgery,* 83(4) Suppl: 13–26.

Perkins, R. Seddon, M. and Effective Practice Informatics and Quality (EPIQ) (2006). 'Quality improvement in New Zealand healthcare. Part 5: measurement for monitoring and controlling performance – the quest for external accountability.' *New Zealand Medical Journal,* 119(1241): U2149.

Petersen, LA. Woodard, LD. Urech, T. Daw, C. Sookanan, S (2006). 'Does pay-for-performance improve the quality of health care?' *Annals of Internal Medicine,* 145(4): 265–272.

Quan, H. Parsons, GA. Ghali, WA (2002). 'Validity of information on comorbidity derived from ICD-9-CM administrative data.' *Medical Care,* 40(8): 675–685.

Roland, M (2004). 'Linking physicians' pay to the quality of care – a major experiment in the United Kingdom.' *New England Journal of Medicine,* 351(14): 1448–1454.

Romano, PS. Chan, BK. Schembri, ME. Rainwater, JA (2002). 'Can administrative data be used to compare postoperative complication rates across hospitals?' *Medical Care,* 40(10): 856–867.

Rosenthal, MB (2007). 'Nonpayment for performance? Medicare's new reimbursement rule.' *New England Journal of Medicine,* 357(16): 1573–1575.

Scott, IA (2007). 'Pay for performance in health care: strategic issues for Australian experiments.' *Medical Journal of Australia,* 187(1): 31–35.

Shahian, DM. Edwards, FH. Ferraris, VA. Haan, CK. Rich, JB. Normand, SL. DeLong, ER. O'Brien, SM. Shewan, CM. Dokholyan, RS. Peterson, ED (2007). 'Quality measurement in adult cardiac surgery: Part 1 – conceptual framework and measure selection.' *Annals of Thoracic Surgery,* 83(Suppl. 4): S3–S12.

Shahian, DM. Silverstein, T. Lovett, AF. Wolf, RE. Normand, SL (2007a). 'Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards.' *Circulation,* 115(12): 1518–1527.

Shekelle, PG (2009). Public performance reporting on quality information. In: Smith, PC. Mossialos, E. Papanicolas, I. Leatherman, S (eds.). *Performance measurement for health system improvement: experiences, challenges and prospects*. Cambridge: Cambridge University Press.

Simborg, DW (1981). 'DRG creep: a new hospital-acquired disease.' *New England Journal of Medicine*, 304(26): 1602–1604.

Terris, DD. Aron, DC (2009). Attribution and causality in health-care performance measurement.' In: Smith, PC. Mossialos, E. Papanicolas, I. Leatherman, S (eds.). *Performance measurement for health system improvement: experiences, challenges and prospects*. Cambridge: Cambridge University Press.

US Department of Health and Human Services (2007). 'Medicare program: changes to the hospital inpatient prospective payment systems and fiscal year 2008 rates.' *Federal Register,* Vol. 72: Sections 47379–47428.

Velasco-Garrido, M. Borowitz, M. Øvretveit, J. Busse, R (2005). Purchasing for quality of care. In: Figueras, J. Robinson, R. Jakubowski, E (eds.). *Purchasing to improve health systems performance*. Berkshire, UK: Open University Press.

WHO (2001). *International classification of functioning, disability and health: ICF*. Geneva: World Health Organization.

Zhan, C. Elixhauser, A. Friedman, B. Houchens, R. Chiang, YP (2007). 'Modifying DRG-PPS to include only diagnoses present on admission: financial implications and challenges.' *Medical Care,* 45(4): 288–291.

3.2 *Clinical surveillance and patient safety*

OLIVIA GRIGG, DAVID SPIEGELHALTER

Introduction

Clinical surveillance is the routine collection of clinical data in order to detect and further analyse unusual health outcomes that may arise from a special cause. As in the closely related subject area of statistical surveillance, the aim is typically to isolate and understand special causes so that adverse outcomes may be prevented. Clinical surveillance is a way of providing appropriate and timely information to health decision-makers to guide their choice of resource allocation and hence improve the delivery of health care.

In order to detect unusual data points, first it is important to take account of the measurable factors that are known to affect the distribution and size of the data. Factors typically of key importance in clinical surveillance are discussed in the first section of this chapter. These include important aspects of clinical surveillance data that affect and govern analysis, including patient heterogeneity; the essential size of health-care facilities; and the dimensionality of the data. Given these essential factors, various statistical surveillance tools might be implemented. Statistical control chart options for surveillance are considered, keeping in mind the desirable characteristics of control charts – utility, simplicity, optimality and verity. A variety of such tools are discussed via example data, with an emphasis on graphical display and desirable characteristics. The graphs presented are based on data relating to cardiac surgery performed by a group of surgeons in a single cardiothoracic unit, and on data relating to the practice of Harold Shipman over the period 1987–1998.

Clinical surveillance: important aspects of the data

We consider four aspects of clinical surveillance data in particular: (i) patient demographics; (ii) throughput of health-care facilities or

providers; (iii) overdispersion in measured quality indicators; and (iv) dimensionality of the data collected.

Patient demographics

Patients arrive at health-care facilities in varying states of health. Any differences observed in the quality of care that health-care facilities provide might be explained in part by variations in the demography of their catchment populations. Aspects of the demography affecting the burden of the health-care facilities (particularly patient mix and the essential size of the community they serve) might affect measured indicators of quality of care. The relationship between these demographic factors and quality of care indicators might be described through a statistical model of risk (see, for example, Cook et al. 2003; Steiner et al. 2000) that can be used as a guide to express the functional state of health-care facilities and systems. Such a model would predict or describe patients' care experience for a variety of patient categories. Future measurements of quality of care indicators could be compared to the risk model that is updated as and when required.

Alternatively, direct stratified standardization might be applied prospectively to panel or multistream data collected over a group of health-care facilities or providers (Grigg et al. 2009; Rossi et al. 1999). This type of adjustment at each time period for the mix and volume of patients across providers allows for surveillance of change within and between providers, but not overall. The latter requires a well-defined baseline against which to check for change, perhaps in the form of a risk model.

Throughput of providers and health-care facilities

Quality of care measures or indicators that are based on rates or counts require an appropriate denominator that represents, or captures some aspect of, the throughput of the health-care facility. In some circumstances this denominator might be viewed as a surrogate for the absolute size of a health-care facility. In cross-sectional comparisons (across health-care facilities or providers) of measures of quality based on rates or counts, the denominator may vary. If there is a common underlying true rate, measured rates associated with larger denominators should vary less about that rate than those associated with smaller

denominators. Hence, in charts that plot the measured rates against an appropriate denominator the points tend to form the shape of a funnel (Spiegelhalter 2005; Vandenbroucke 1988).

Overdispersion amongst outcomes

Unmeasured case-mix or demographic factors may produce overdispersion amongst quality indicators measured across health-care facilities. In such cases the statistical model that relates those factors to quality of care may not apply precisely at all time points to all of the facilities (Aylin et al. 2003; Marshall et al. 2004). Given the risk model, the variability in outcomes may be substantially higher than that expected from chance alone and the excess not explainable by the presence of a few outlying points. This overdispersion (or general lack of fit to the whole population of health-care facilities) might be expressed through hierarchical models that would allow for slack in the fit of the risk model, or in standardized risk measures across facilities (Daniels & Gatsonis 1999; Grigg et al. 2009; Ohlssen et al. 2007). Time-dependent hierarchical models might also allow for flexibility or evolution of the risk model over time (Berliner 1996; West & Harrison 1997).

Dimensionality of the data

The higher the number of health-care facilities or providers that are compared then the greater the potential for false positive results or significant departures from the model describing the normal functional state of the facilities. This is due to the assumed inherent randomness in the system. The potential for false positive results of significance also increases if many quality of care indicators are measured and monitored repeatedly over time. Possible approaches for handling the multivariate nature of the monitoring problem and controlling the multiplicity of false positives include:

- describing the system as a multivariate object and employing multivariate control charts in which signals generally relate only to the system as a whole and require diagnosis to establish any smaller scale causes (Jackson 1985; Lowry & Montgomery 1995);

- employing univariate control charts, mapping the univariate chart statistics to a reference scale and then applying a multiplicity controlling procedure to the multivariate set of mapped values (Benjamini & Kling 1999; Grigg et al. 2009);
- comparing potentially extreme observed chart statistic values to a large population of chart statistic values simulated under null conditions and checking whether those observed values still appear significant (Kulldorf et al. 2007).

Statistical chart options

A wide range of charting tools has been suggested for surveillance of health measures over time, largely adapted from the industrial quality-control context (Woodall 2006). We now describe some of these charting tools, with an emphasis on desirable characteristics.

The charts illustrated include the Shewhart chart; scan statistic, moving average (MA), exponentially weighted moving average (EWMA), sets method, cumulative O – E, cumulative sum (CUSUM) and maximized CUSUM. We illustrate all but the last method using data relating to a group of seven cardiac surgeons in a single cardiac unit. We illustrate the maximized CUSUM using data relating to the practice of the late Harold Shipman, general practitioner and convicted murderer, over the period 1987 to 1998. We consider that the desirable characteristics of a charting tool are:

- *Utility:* ease of interpretation of the graphic; intuitiveness of presentation from a general user's point of view.
- *Simplicity* of the mathematics behind the chart (regarding the chart algorithm calculation of operating characteristics; and calculation of bands, bounds or limits).
- *Responsiveness* (under any circumstances) to important and definable but perhaps subtle changes, where these can be discriminated from false alarms.
- *Verity:* graphical effectiveness and ability to give a close and true description of the process.

It is well known that the CUSUM and EWMA rate highly on responsiveness and the Shewhart chart rates highly on simplicity. Utility and

verity are more subjective and therefore it is difficult to say which of the charts, if any, rate highly on these. However, we will attempt to provide some assessment.

Example data: cardiac surgery

Fig. 3.2.1 is a plot (by surgeon) of outcomes adjusted for patient pre-operative risk against operation number. The operation number is the time-ordered operation number and is measured collectively over operations performed by any one of the seven surgeons. The outcomes are coded so that 0 ≡ patient survival past thirty days following surgery, 1 ≡ death of a patient within thirty days.

The outcomes are adjusted by the use of a model calibrated on the first 2218 operations that relates the patient Parsonnet score to the probability of not surviving beyond thirty days (Parsonnet et al. 1989; Steiner et al. 2000). The adjustment leads to data of the form *observed* – *expected* + *baseline*, where the baseline is the mean thirty-day mortality rate in the calibration dataset (= 0.064, given 142 deaths) and the expected outcome is calculated from the risk model. For example, the adjusted outcome for a patient with an expected risk of 0.15 is 1 - 0.15+0.064 = 0.914 if he/she does not survive beyond thirty days following surgery but - 0.15 + 0.064 = -0.086 if she/he does. If the model described predicts patient risk well, the adjustment should increase the comparability of the outcomes of operations performed on differing types of patients.

The adjusted outcomes relating to operations performed by each of the seven surgeons are plotted in grey (Fig. 3.2.1). Points falling at or below zero on the risk-adjusted outcomes scale correspond to patients who survived beyond thirty days; points falling above correspond to those who did not. A smooth mean of the adjusted outcomes is plotted in black (calculated over non-overlapping windows of time, 250 operations in duration) and can be compared to the mean thirty-day mortality rate of 0.064 from the calibration data. These mean adjusted outcomes are plotted on a finer scale in Fig. 3.2.3, with pointwise significance bands or p-value lines (see below).

The extremity of a patient's pre-operative condition is indicated by the extent to which the grey adjusted outcomes in Fig. 3.2.1 fall from the original data values of 0 and 1. For Surgeon 1, a large density of points fall below their original data values of 0 and 1 but Fig.

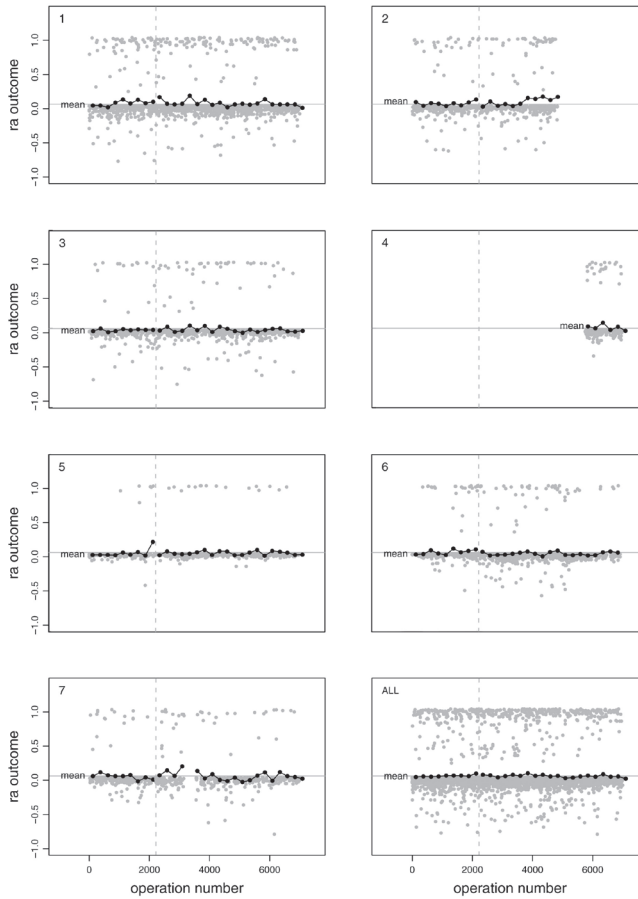

Fig. 3.2.1 Risk-adjusted outcomes (adjusted thirty-day mortality, given patient Parsonnet score) relating to operations performed in a cardiac unit in which there are seven surgeons. First 2218 data are calibration data.

3.2.2 shows that this is because this surgeon consistently receives and treats high-risk patients (with high Parsonnet scores). In contrast, the adjusted outcomes for Surgeon 5 are closer to the original data values as this surgeon consistently receives and treats lower risk patients (see Fig. 3.2.2).

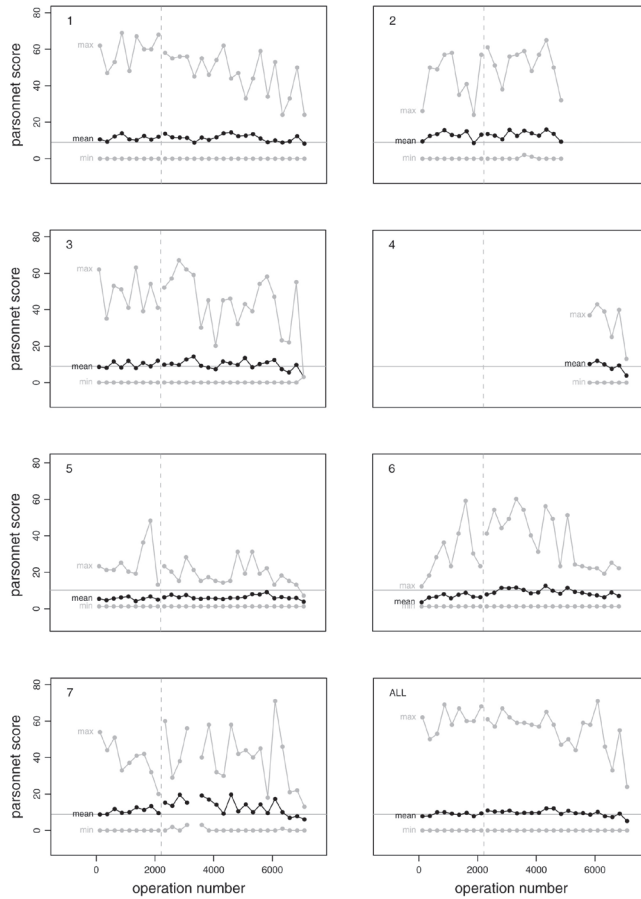

Fig. 3.2.2 Parsonnet score of patients treated in a cardiac unit.

Shewhart charts, scan statistics and MAs

Shewhart charts (Shewhart 1931) plot each individual data point or groups of data points if the data are highly discrete e.g. binary data. Dependent on the size of these groups, the charts can provide quite smooth estimates of the current underlying risk. The charts will only be able to detect departures from baseline risk that affect groups at least as big as those comprising the data-points. A plotted value that falls outside a sufficiently small significance band is evidence of departure from the baseline risk model.

Fig. 3.2.3 is a plot by surgeon of the mean risk-adjusted outcome over disjoint windows of 250 operations performed by all of the surgeons. The plotted binomial significance bands are similar to bands marked on funnel plots (Spiegelhalter 2005) in that they change according to the number of operations performed by an individual surgeon in each window. This number is essentially the denominator used to calculate the bands. If one surgeon performed many of the operations in a window then their chart for that window would have narrow bands. It can be seen that Surgeons 1 and 6 generally perform the most operations out of the group, since the significance bands on charts 1 and 6 are tighter than those on the other charts. The bands on the chart of mean risk-adjusted outcome for all surgeons do not change over time, except for the final incomplete window of 54 observations, as they are based on a constant denominator of 250.

The charts in Fig. 3.2.3 can be viewed as types of Shewhart chart (Shewhart 1931), where the control limits or significance bands are adjusted for the volume of patients treated by a surgeon in each window of time. Equivalent risk-adjusted Shewhart charts could be drawn by plotting the mean of the original data values and adjusting the significance bands for patient case-mix, or Parsonnet score, as well as the denominator (Cook et al. 2003; Grigg & Farewell, 2004).

The charts in Fig. 3.2.3 are also related to the scan statistic method (Ismail et al. 2003). This method retrospectively detects areas or clusters of lack of agreement with the risk model by conditioning on there being such a cluster and then locating it. This method indicates that the most concentrated area of lack of agreement with the model is around operation number 3500 (in an upwards direction) for Surgeon 1 and around operation number 4500 (in a downwards direction) for Surgeon 6. For the group of surgeons as a whole, the method indicates that the most concentrated areas of lack of agreement with the risk model are around operation numbers 4000 (upwards) and 5000 (downwards).

For scan statistic methods it is more typical to scan the data via a moving window (moving one observation at a time) than to scan over neighbouring and non-overlapping windows. The charts in Fig. 3.2.4 can be viewed as performing the former, as they plot each surgeon's MAs for sets of thirty-five adjusted outcomes. The MA is updated for each surgeon for every operation and so is updated more often for those who receive patients regularly (e.g. Surgeon 1) than for those

who receive patients less frequently (e.g. Surgeon 5). The MAs can be compared against significance bands calculated in the same way as those in Fig. 3.2.3, but the denominator remains at a constant value of thirty-five. As might be expected, in any particular chart of Fig. 3.2.4, the frequency of evidence indicating lack of agreement with the risk model appears to be related to how frequently the surgeon operates. This can be seen on the chart for all surgeons, which is the most volatile and spiky. In theory the mathematical design of these charts is simple – plotting a summary statistic of groups of data points in which points within groups carry equal weight. The charts should rate quite highly on utility, verity and responsiveness if the aims of the design are met, i.e. the summary statistic summarizes the original data points well and the chosen group size is appropriate. However, the constraint of equal weightings of data points may limit the verity of the charts and their simplicity may be affected if the form of summary statistic and the size of groups of the charts are treated as parameters to be optimized.

EWMAs

Similarly to the charts described immediately above, the EWMA chart (Roberts 1959) provides a smoothed estimate of the current underlying risk but uses all past data since initialization of the chart. Fig. 3.2.5 shows plots of EWMAs (by surgeon) of the risk-adjusted outcomes, with accompanying credible intervals for the mean thirty-day mortality rate at operation number t associated with surgeon j, μ_{tj}, as it evolves from the baseline value μ_0 calculated across all surgeons in the calibration dataset. Any given plotted EWMA value on a particular surgeon's chart is a weighted average of all previous adjusted outcomes for that surgeon. The weights decay geometrically by a factor $\kappa = 0.988$ so that less recent outcomes are given less weight than recent outcomes. The value of κ was chosen so as to minimize the mean squared error of prediction of patient thirty-day mortality in the calibration dataset. The EWMA plotted at operation number t performed by surgeon j can be written as:

$$\omega_{0j} = \mu_0 \qquad (1)$$
$$\omega_{tj} = \kappa\omega_{t\text{-}1,j} + (1 - \kappa)Y_{tj}, \quad t = 1,2, \ldots \qquad j = 1,2, \ldots, 7.$$

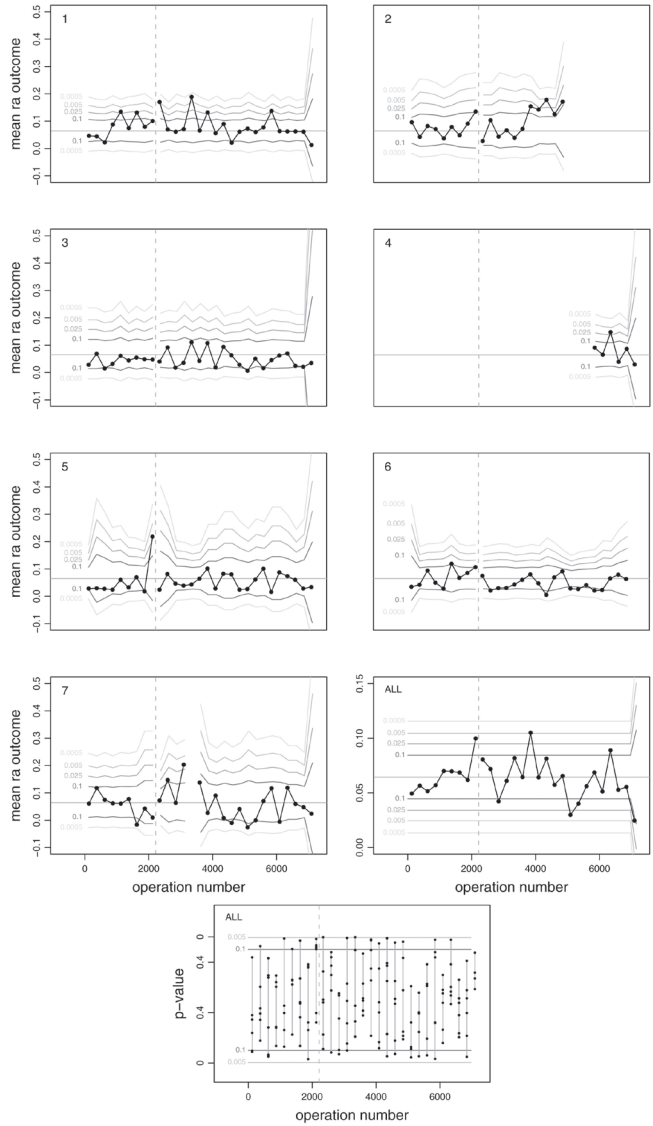

Fig. 3.2.3 Mean risk-adjusted outcome over disjoint windows of 250 operations, where operations are by any of seven surgeons in a cardiac unit. Bands plotted are binomial percentiles around the mean patient 30-day mortality rate from the calibration data ($\mu_0 = 0.064$), where the denominator is the number of operations by a surgeon in a given window. Gaps in the series other than at the dashed division line correspond to periods of inactivity for a surgeon.

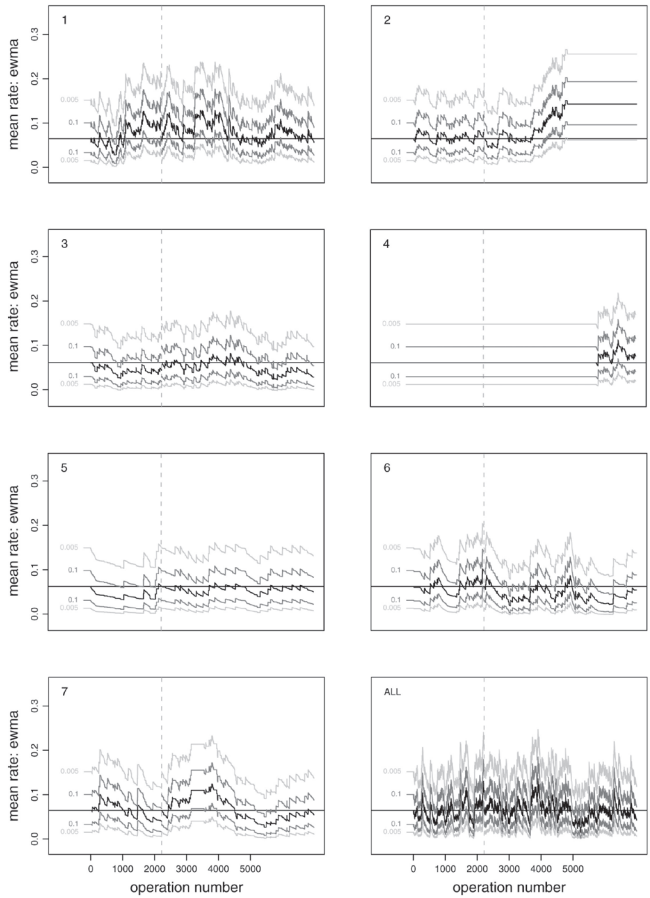

Fig. 3.2.4 Moving average (MA) of risk-adjusted outcomes over overlapping windows of 35 operations by a particular surgeon from a cardiac unit of seven surgeons. Bands plotted are binomial percentiles around the mean patient 30-day mortality rate from the calibration data ($\mu_0 = 0.064$), where the denominator is 35.

where $\mu_0 = 0.064$ is the mean thirty-day mortality rate in the calibration dataset and $Y_{tj} = O_{tj} - E_{tj} + \mu_0$ is the adjusted observation at time t relating to surgeon j.

Equivalently, we can write:

$$\omega_{0j} = \mu_0 \qquad (2)$$
$$\omega_{tj} = \omega_{t\text{-}1,j} + (1 - \kappa)(O_{tj} - E_{tj}),\ t = 1, 2, \ldots\ j = 1, 2, \ldots, 7.$$

To calculate the credible intervals it is assumed that a distribution for the mean patient thirty-day mortality rate at operation number t and relating to surgeon j, μ_{tj}, can be described as beta with mean given by the EWMA estimate ω_{tj} and precision given by $(1 - \kappa)^{-1} = 83.3$. Grigg & Spiegelhalter (2007) provide further discussion about these intervals and the risk-adjusted EWMA.

The charts in Figs. 3.2.2–3.2.4 have significance bands or control lines drawn around a calibrated mean but in the EWMA drawn here bounds are placed around the chart statistic. The bounds placed describe uncertainty in the estimate of the current underlying risk. Despite the change of emphasis, lack of agreement with the risk model on any particular chart can still be investigated by checking the extent to which the credible bounds around the EWMA statistic cross the baseline mean patient thirty-day mortality rate, $\mu_0 = 0.064$. A lack of agreement with the risk model is indicated if μ_0 falls far into the tails of the plotted distribution for μ_{tj}.

As seen in Fig. 3.2.5, the outermost credible bounds (at a p-value of ±0.0005) drawn for the distribution of the mean patient thirty-day mortality rate in relation to surgeon j remain mostly below a rate of 0.2 on all the charts. EWMA charts might be considered to have a more complex mathematical design than Shewhart charts as the weighting of data points is not necessarily equal. The chart statistic includes all past data since the start of the chart. This should improve the verity of the estimation of the true current underlying risk but may reduce the responsiveness if the weighting parameter is not well-tuned. The placement of bounds around the chart statistic may affect the utility of the chart, dependent on the user, but again this should improve the verity of estimating the true current underlying risk.

Sets method

The sets method (Chen 1978) measures the number of outcomes occurring between outcomes classified as events. Typically, a signal is given if the set size is less than a value T on n successive occasions, where T and n can be tuned so that the chart is geared towards testing for a specific shift in rate (Gallus et al. 1986). For example, a signal might be given if there were three non-survivors within the space of twenty operations.

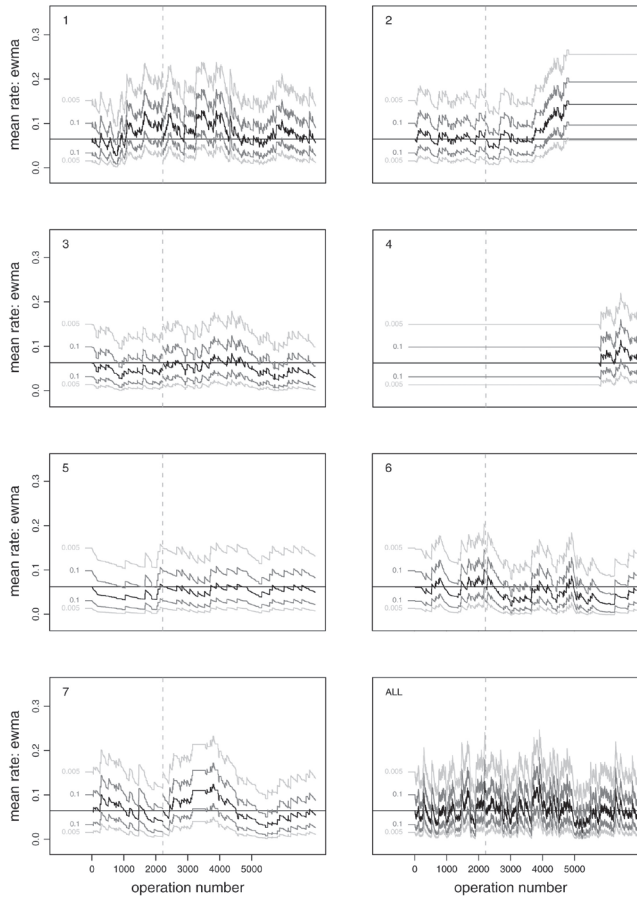

Fig. 3.2.5 Exponentially weighted moving average (EWMA) of risk-adjusted outcomes of surgery by a particular surgeon from a cardiac unit of seven surgeons. Less recent outcomes are given less weight than recent outcomes, by a factor of $k = 0.988$. The EWMA and accompanying bands give a running estimate by surgeon of the mean patient 30-day mortality rate and uncertainty associated with that estimate.

Fig. 3.2.6 shows risk-adjusted sets charts by surgeon, where the adjusted number of operations between surgical outcomes coded 1 (patient survives less than 30 days following surgery) is plotted against operation number. As discussed by Grigg & Farewell (2004b), the adjustment of the accruing set size at each observation is such that higher-than-average risk patients contribute more to the set size than

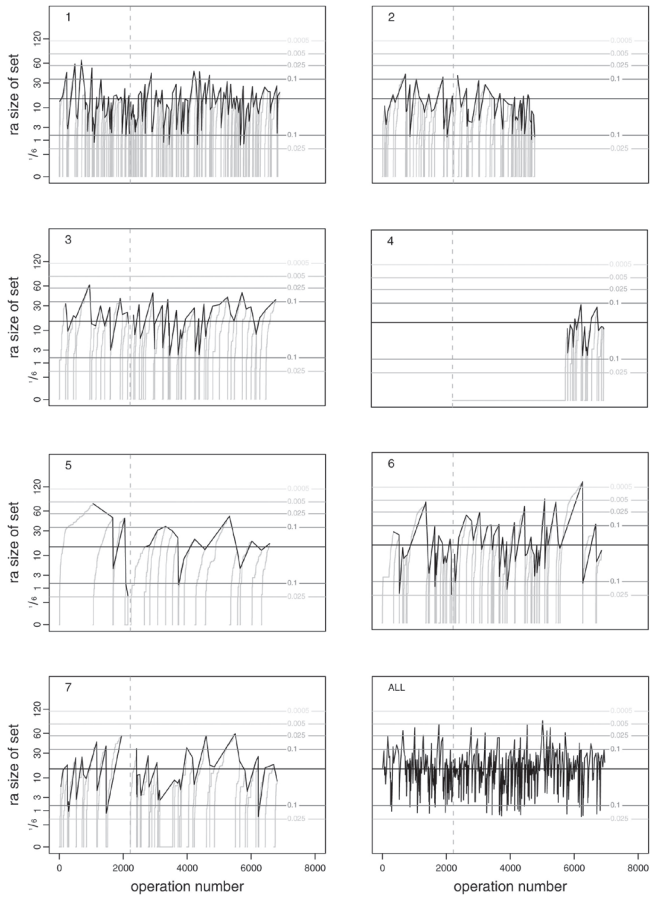

Fig. 3.2.6 Risk-adjusted set size, or adjusted number of operations between outcomes of 1 (where a patient survives less than 30 days following surgery), associated with surgery by a particular surgeon from a cardiac unit of seven surgeons. Bands plotted are geometric percentiles based on the mean patient 30-day mortality rate from the calibration data ($\mu_0 = 0.064$).

those with average risk (risk equal to the baseline risk, $\mu_0 = 0.064$) and lower risk patients contribute less than those with average risk.

The accruing adjusted set size for surgeon j at operation number t, which resets to zero when the observed outcome from the previous operation $O_{t-1,j}$ equals 1, can be written as:

$$S_{0j}=0 \qquad (3)$$

$$S_{tj}=\left(S_{t-1,j}+\frac{E_{tj}}{\mu_0}\right)(1-0_{t-1,j})+\left(\frac{E_{tj}}{\mu_0}\right)0_{t-1,j},\ t=1,2,\ \ldots\ j=1,2,\ \ldots,7.$$

where E_{tj} is the expected outcome at operation number t performed by surgeon j and is calculated from the risk model. This accruing set size is plotted in grey on the charts in Fig. 3.2.6. The absolute set sizes are joined up in black, at the points where the observed outcome O_{tj} equals 1. The significance bands plotted are geometric and calibrated about the baseline expected set size calculated from the first 2218 observations, $1/\mu_0 = 15.63$.

A noteworthy result from these charts is the very large adjusted set size of 132 recorded on the chart for Surgeon 6 at around operation number 6000. This magnitude of set size is interpretable as equivalent to a run of over 132 operations performed on baseline risk patients where those patients all survive beyond 30 days following surgery.

The plots drawn in Fig. 3.2.6 might be viewed as more complex than Shewhart charts of the number of outcomes between events, since the accruing risk-adjusted set size is also plotted. As with runs rules on Shewhart charts (Western Electric Company 1984), a more complex stopping rule may improve the responsiveness, but affect utility. The transformation (Nelson 1994) of the y-axis in Fig. 3.2.6 is intended to ensure that the verity or utility of the charts should not be affected by the fact that they plot time between event data rather than rate data.

Cumulative O – E and CUSUM charts

The cumulative charts described here accumulate measures of departure from the baseline risk model, where the accumulation is either over all outcomes since the start of the chart or is adaptive according to the current value of the chart statistic.

The charts in Fig. 3.2.7 show each surgeon's cumulative sum of *observed-expected* outcomes from surgery (cumulative O – E) where the expected counts are calculated using the risk model relating patient thirty-day mortality to Parsonnet score. This type of chart has also been called a variable life-adjusted display (VLAD) (Lovegrove et al. 1997; Lovegrove et al. 1999) and a cumulative risk-adjusted mortality chart (CRAM) (Poloniecki et al. 1998). The cumulative O – E chart statistic at operation number t relating to surgeon j can be written as:

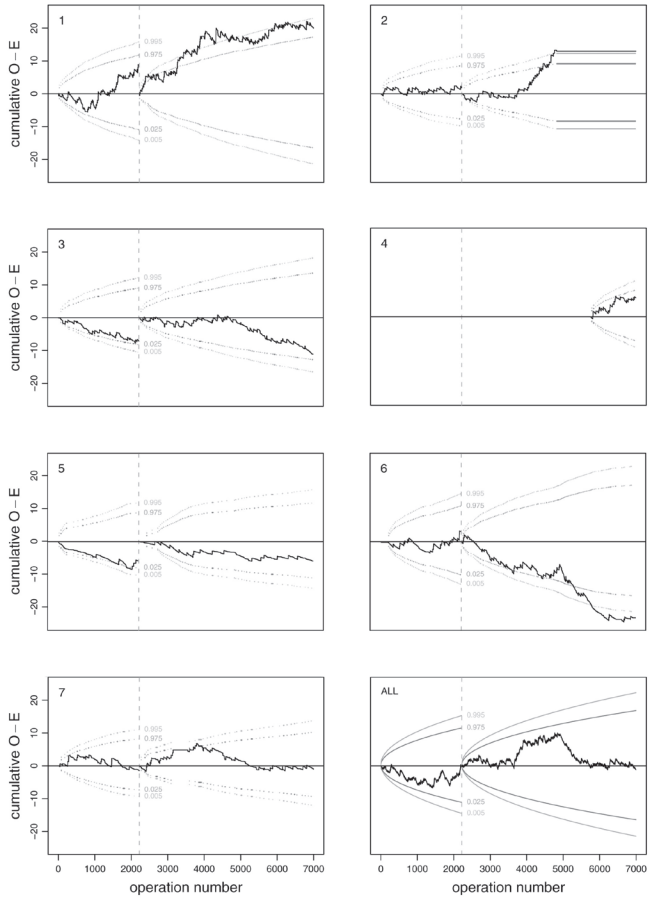

Fig. 3.2.7 Cumulative sum of observed outcome, from an operation by a particular surgeon from a cardiac unit of seven surgeons, minus the value predicted by the risk model given patient Parsonnet score. Bands plotted are centered binomial percentiles based on the mean patient 30-day mortality rate from the calibration data ($\mu_0 = 0.064$).

$$V_{0j} = 0 \qquad (4)$$
$$V_{tj} = V_{t\text{-}1,j} + O_{tj} - E_{tj},\ t = 1,2,\ldots\ j = 1,2,\ldots,7$$

The charts display each surgeon's accruing excess patient thirty-day mortality above that predicted by the risk model given patient pre-operative risk, where this is assumed to be described by patient Parsonnet score. The measure accrued is simple (except perhaps in its

reliance on the accuracy of the risk model) but the charts may be easy to misinterpret. For example, Surgeon 1's chart reaches an excess of 20 patient mortalities above that predicted by the risk model at around operation number 4000. However, the chart retains any past excess and therefore indicates that this excess continues at approximately the same level. Given the accuracy of the risk model, information about a surgeon's current operative performance is mostly contained in the gradient of these charts. This is indicated by the increase in the significance bands on the charts each time a surgeon operates.

The CUSUM chart (Hawkins & Olwell 1997) is closely related to the cumulative O – E chart. However, it accumulates a function of the observed and expected outcomes that reflects the relative likelihood of the baseline risk model compared to that of an alternative model, given the surgical outcomes observed since the start of the chart. This accmulated measure is an optimal measure of departure (Moustakides 1986) and thus these charts are very responsive to important changes, i.e. movement towards alternative models. The chart maintains sensitivity to departure from the baseline model by accumulating only evidence in favour of the alternative model, otherwise it remains at the balance point (zero).

In Fig. 3.2.8, CUSUM charts on the observed outcomes are plotted by surgeon. The upper half of the chart tests for a doubling in the odds of patient thirty-day mortality; the lower half tests for a halving. The significance bands, or p-value lines, are based on the empirical distribution of CUSUM values simulated under baseline conditions. More discussion on associating CUSUM values with p-values can be found in Benjamini and Kling (1999) and Grigg and Spiegelhalter (2008).

The CUSUM chart statistic at operation number t relating to surgeon j can be written as:

$$C_{0j}=0 \qquad (5)$$

$$C_{tj}=max\left\{0,C_{t-1,j}+log\left[\frac{P(0_{tj}\mid alternative)}{P(0_{tj}\mid baseline)}\right]\right\},\ t=1,2,\ \ldots\ \ j=1,2,\ \ldots,7.$$

If, as in the charts plotted in Fig. 3.2.8, the alternative model specifies a uniform change (R) from the baseline model across patient types of the odds of thirty-day mortality, the CUSUM chart statistic can be written as:

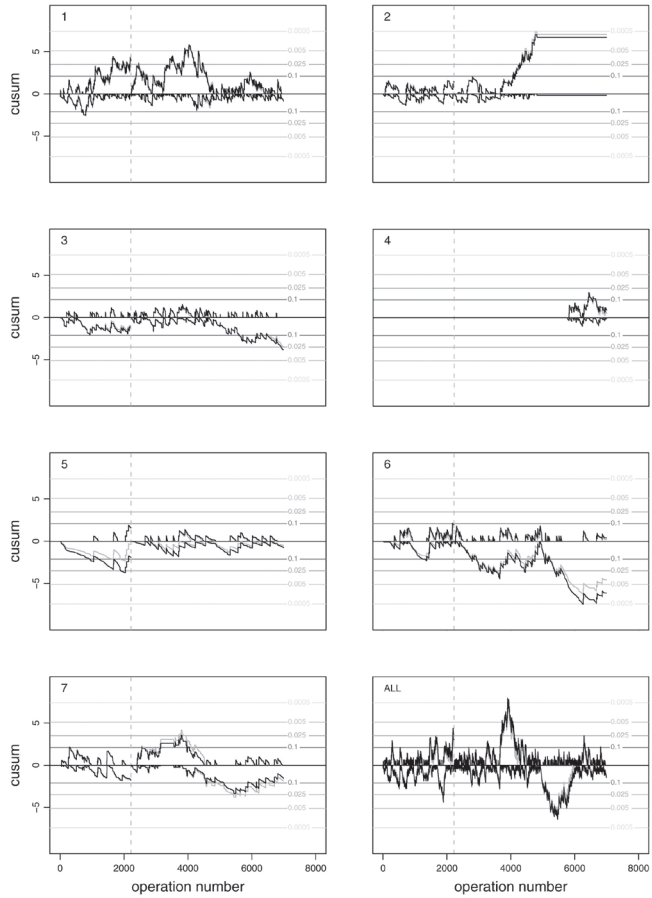

Fig. 3.2.8 Cumulative log-likelihood ratio of outcomes from operations by a particular surgeon from a cardiac unit of seven surgeons, comparing the likelihood of outcomes given the risk model with that given either elevated or decreased risk. Upper chart half is a CUSUM testing for a halving in odds of patient survival past 30 days, lower chart half for a doubling in odds of survival past 30 days.

$$C_{0j}=0 \tag{6}$$

$$C_{tj}=max\left\{0,C_{t-1,j}+log(R)0_{tj}-\left[\frac{log(1-E_{tj}+RE_{tj})}{E_{tj}}\right]E_{tj}\right\},\ t=1,2,\ldots\ j=1,2,\ldots,7.$$

As noted by Grigg et al. (2003), the chart statistic increments are then seen to be of the form $aO - b(E)E$, and hence similar to the $O - E$

form in Fig. 3.2.7. In particular, for $R = 2$ the increments are approximately $(\log 2)O_{tj} - E_{tj}$.

Exact risk-adjusted CUSUMs (Steiner et al. 2000) based on the original outcomes and the full likelihood (given the risk model) are plotted in black in Fig. 3.2.8. CUSUMs based on the adjusted outcomes $O_{tj} - E_{tj} + \mu_0$ and the unconditional likelihood are plotted in grey. These closely follow the exact CUSUMs, thereby illustrating that the likelihood contribution from the adjusted outcomes is approximately equivalent to that from the original outcomes. This point is noted in the section on example data for cardiac surgery and described by Grigg and Spiegelhalter (2007).

The Shewhart chart for all surgeons (Fig. 3.2.3) suggests a lack of agreement with the null model around operation numbers 4000 (in an upwards direction) and 5000 (in a downwards direction). This can also be seen in the CUSUM chart for all surgeons (Fig. 3.2.8) but here the evidence of potential lack of agreement is more pronounced. The CUSUM is known to be responsive but this may be at the expense of simplicity and utility. A maximized CUSUM (see section below) may improve the verity of the chart.

Example data: Harold Shipman

Fig. 3.2.9 is a plot of maximized CUSUM charts by age-sex groupings of patients registered with general practitioner Harold Shipman over the period 1987 to 1998 (Baker 2001; Shipman Inquiry 2004). In 2000, Harold Shipman was convicted for murdering fifteen of his patients but he may have killed two hundred (Baker 2001; Shipman Inquiry 2002 & 2004; Spiegelhalter et al. 2003). The chart statistics in Fig. 3.2.9 are as described by equation 5, except that a vector of CUSUM statistics (rather than a single CUSUM statistic) is plotted on each half of the chart. A Poisson likelihood is adopted as the data are grouped mortality counts; the section on cumulative O – E and CUSUM used the Bernoulli likelihood as the data relate to individual patients. The baseline risk for a particular age-sex category is taken to be the England and Wales standard in any given year, as described in Baker (2001).

Each element of the plotted vector corresponds to a CUSUM comparing a particular alternative model to the baseline risk model. On the upper half of the chart, the alternative ranges from no change

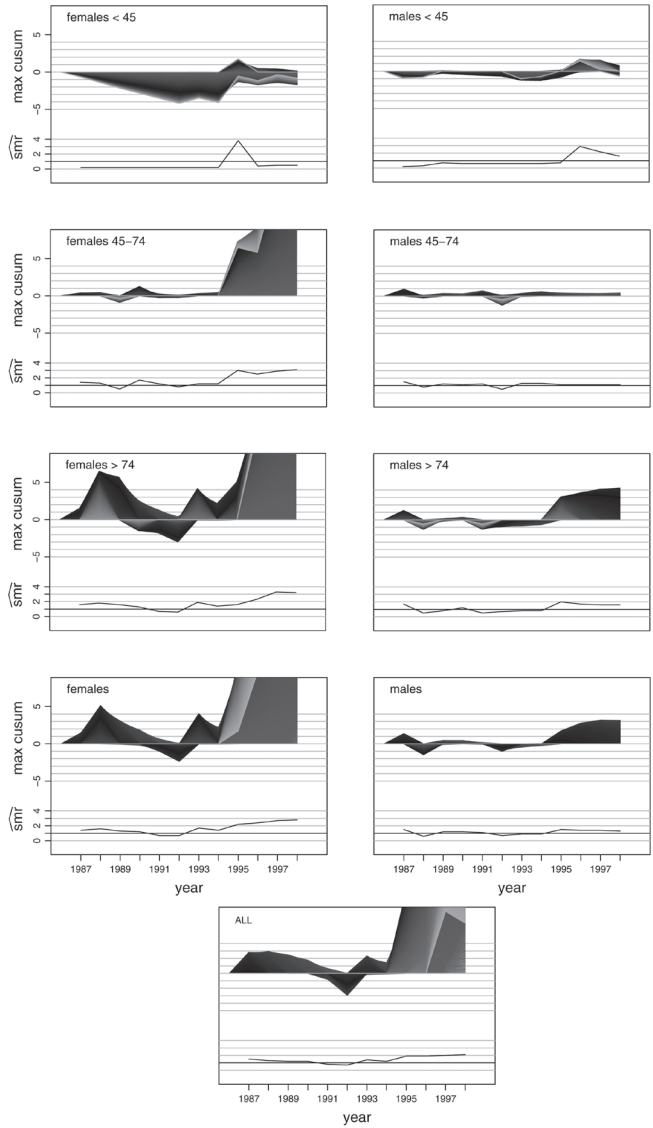

Fig. 3.2.9 Maximised CUSUM of mortality outcomes by age-sex category of patients registered with Harold Shipman over the period 1987–1998, comparing the likelihood of outcomes under the England and Wales standard with that given either elevated or decreased risk. Upper chart half is testing for up to a four-fold increase in patient mortality, lower chart halffor up to a four-fold decrease. The estimated standardised mortality rate (SMR) is given.

in risk to a uniform four-fold increase in patient risk across all age-sex categories. Similarly, on the lower half, the alternative ranges from no change in risk to a uniform four-fold decrease in patient risk.

On each half of the chart the external edge of the block of plotted vectors corresponds to the most extreme value in the vector of CUSUM values at any one time. This may relate to different alternative models over time; the alternative model that they relate to represents the best supported alternative to the baseline model (Lai 1995; Lorden 1971). In this way, the maximized CUSUM gives both the maximized evidence in favour of non-baseline risk models and the specific alternative at any one time that corresponds to the maximized evidence.

The pattern of the chart for females over seventy-four can be seen to dominate the chart for all females as well as the overall chart for all patient categories. The estimated standardized mortality ratio (corresponding to the maximized CUSUM value) on the chart for females over seventy-four increases from 1.5 in 1994 to more than 3 in the years 1997 to 1998. From 1995 there is strong evidence of increasing departure from the baseline risk model. A similar increase in estimated SMR is seen on the chart for females aged between forty-five and seventy-four. The increase is mirrored but dampened in the chart for all females and dampened further in the chart for all patients. This dampening is due to information added from the other charts and illustrates why comparisons of outcomes across different aspects of a dataset are hampered by the 'curse of dimensionality' (Bellman 1957).

Conclusions

We have described a selection of statistical control charts that could (individually or in combination) form a basis for clinical surveillance. The charts described include: fixed window methods, e.g. Shewhart, scan statistic and MA charts; continuous window methods, e.g. EWMA and O – E charts; and adaptive window methods e.g. sets method, CUSUM and maximized CUSUM. The charts are graphically illustrated through some example data which include cardiac surgery outcomes, from operations performed in the period 1992-1998 by a group of surgeons in a single cardiothoracic unit, and mortality outcomes of patients registered with Harold Shipman in the period 1987–1998.

We have suggested some desirable characteristics (utility, simplicity, responsiveness, verity) that might be considered when deciding which

charts to include in a clinical surveillance system. Our discussion indicates that simpler charts such as the fixed window methods are likely to have better utility but may compromise responsiveness and verity. Verity should be high if a chart gives a running estimate with bounds of the parameter of interest, where these bounds reflect uncertainty surrounding the estimate. The maximized CUSUM can provide such an estimate and is known to be responsive. The EWMA is similarly responsive but may be simpler than the maximized CUSUM, as the chart gives a direct running estimate.

Each of the charts has a variety of characteristics that may be comparable but we recommend the use of a combination of charts, with simpler charts in the foreground. Further, we recommend that any practical application of the charts should be embedded in a structured system for investigating any signals that might be detected.

References

Aylin, P. Best, N. Bottle, A. Marshall, C (2003). 'Following Shipman: a pilot system for monitoring mortality rates in primary care.' *Lancet*, 362(9382): 485–491.

Baker, R (2001). *Harold Shipman's clinical practice 1974–1998: a review commissioned by the Chief Medical Officer*. London: Stationary Office Books.

Bellman, RE (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Benjamini, Y. Kling, Y (1999). *A look at statistical process control through the p-values*. Tel Aviv: Tel Aviv University (Tech. rept. RP-SOR-99-08 http://www.math.tau.ac.il/~ybenja/KlingW.html).

Berliner, L (1996). *Hierarchical Bayesian time series models*. Cite Seer X – Scientific Literature Digital Library and Search Engine. (http://citeseer.ist.psu.edu/121112.html).

Chen, R (1978). 'A surveillance system for congenital malformations.' *Journal of the American Statistical Association*, 73: 323–327.

Cook, DA. Steiner, SH. Farewell, VT. Morton, AP (2003). 'Monitoring the evolutionary process of quality: risk-adjusted charting to track outcomes in intensive care.' *Critical Care Medicine*, 31(6): 1676–1682.

Daniels, MJ. Gatsonis, C (1999). 'Hierarchical generalized linear models in the analysis of variations in health care utilization.' *Journal of the American Statistical Association*, 94(445): 29–42.

Gallus, G. Mandelli, C. Marchi, M. Radaelli, G (1986). 'On surveillance methods for congenital malformations.' *Statistics in Medicine*, 5(6): 565–571.

Grigg, O. Farewell, V (2004). 'An overview of risk-adjusted charts.' *Journal of the Royal Statistical Society: Series A*, 167(3): 523–539.

Grigg, OA. Farewell, VT (2004a). 'A risk-adjusted sets method for monitoring adverse medical outcomes.' *Statistics in Medicine*, 23(10): 1593–1602.

Grigg, OA. Spiegelhalter, DJ (2007). 'A simple risk-adjusted exponentially weighted moving average.' *Journal of the American Statistical Association*, 102(477): 140–152.

Grigg, OA. Spiegelhalter, DJ (2008). 'An empirical approximation to the null unbounded steady-state distribution of the cumulative sum statistic.' *Technometrics,* 50(4): 501–511.

Grigg, OA. Farewell, VT. Spiegelhalter, DJ (2003). 'Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts.' *Statistical Methods in Medical Research*, 12(2): 147–170.

Grigg, OA. Spiegelhalter, DJ. Jones, HE (2009). 'Local and marginal control charts applied to methicillin resistant Staphylococcus aureus bacteraemia reports in UK acute NHS Trusts.' *Journal of the Royal Statistical Society: Series A,* 172(1): 49–66.

Hawkins, DM. Olwell, DH (1997). *Cumulative sum charts and charting for quality improvement*. New York: Springer.

Ismail, NA. Pettit, AN. Webster, RA (2003). ''Online' monitoring and retrospective analysis of hospital outcomes based on a scan statistic.' *Statistics in Medicine*, 22(18): 2861–2876.

Jackson, JE (1985). 'Multivariate quality control.' *Communication Statistics – Theory and Methods*, 14(11): 2657–2688.

Kulldorff, M. Mostashari, F. Duczmal, L. Yih, WK. Kleinman, K. Platt, R (2007). 'Multivariate scan statistics for disease surveillance.' *Statistics in Medicine*, 26(8): 1824–1833.

Lai, TL (1995). 'Sequential changepoint detection in quality control and dynamical systems.' *Journal of the Royal Statistical Society: Series B*, 57(4): 613–658.

Lorden, G (1971). 'Procedures for reacting to a change in distribution.' *Annals of Mathematical Statistics*, 42(6): 1897–1908.

Lovegrove, J. Sherlaw-Johnson, C. Valencia, O. Treasure, T. Gallivan, S (1999). 'Monitoring the performance of cardiac surgeons.' *Journal of the Operational Research Society*, 50(7): 684–689.

Lovegrove, J. Valencia, O. Treasure, T. Sherlaw-Johnson, C. Gallivan, S (1997). 'Monitoring the results of cardiac surgery by variable life-adjusted display.' *Lancet*, 350(9085): 1128–1130.

Lowry, CA. Montgomery, DC (1995). 'A review of multivariate control charts.' *IIE Transactions*, 27: 800–810.

Marshall, C. Best, N. Bottle, A. Aylin, P (2004). 'Statistical issues in the prospective monitoring of health outcomes across multiple units.' *Journal of the Royal Statistical Society: Series A*, 167(3): 541–559.

Moustakides, GV (1986). 'Optimal stopping times for detecting changes in distributions.' *Annals of Statistics*, 14(4): 1379–1387.

Nelson, LS (1994). 'A control chart for parts-per-million nonconforming items.' *Journal of Quality Technology*, 26(3): 239–240.

Ohlssen, D. Sharples, L. Spiegelhalter, D (2007). 'A hierarchical modelling framework for identifying unusual performance in health care providers.' *Journal of the Royal Statistical Society: Series A*, 170(4): 865–890.

Page, ES (1954). 'Continuous inspection schemes.' *Biometrika*, 41(1–2): 100–115.

Parsonnet, V. Dean, D. Bernstein, AD (1989). 'A method of uniform stratification of risks for evaluating the results of surgery in acquired adult heart disease.' *Circulation*, 779(1): 1–12.

Poloniecki, J. Valencia, O. Littlejohns, P (1998). 'Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery.' *British Medical Journal*, 316(7146): 1697–1700.

Roberts, SW (1959). 'Control chart tests based on geometric moving averages.' *Technometrics*, 42(1): 239–250.

Rossi, G. Lampugnani, L. Marchi, M (1999). 'An approximate CUSUM procedure for surveillance of health events.' *Statistics in Medicine*, 18(16): 2111–2122.

Shewhart, WA (1931). *Economic control of quality of manufactured product*. New York: Van Nostrand.

Shipman Inquiry (2002). *Shipman Inquiry: First Report*. London, UK: HMSO.

Shipman Inquiry (2004). *Shipman Inquiry Fifth Report - Safeguarding patients: lessons from the past, proposals for the future*. London, UK: HMSO (http://www.the-shipman-inquiry.org.uk/fifthreport.asp).

Spiegelhalter, DJ (2005). 'Problems in assessing rates of infection with methicillin resistant *Staphylococcus aureus*.' *British Medical Journal*, 331(7523):1013–1015.

Spiegelhalter, DJ. Grigg, OAJ. Kinsman, R. Treasure, T (2003). 'Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery.' *International Journal for Quality in Health Care*, 15(1): 7–13.

Steiner, SH. Cook, RJ. Farewell, VT. Treasure, T (2000). 'Monitoring surgical performance using risk-adjusted cumulative sum charts.' *Biostatistics*, 1(4): 441–452.

Vandenbroucke, JP (1988). 'Passive smoking and lung cancer: a publication bias?' *British Medical Journal*, 296(6619): 319–392.

West, M. Harrison, J (1997). *Bayesian forecasting and dynamic models. Second edition*. New York: Springer-Verlag.

Western Electric Company (1984). *Statistical quality control handbook*. Texas: AT & T Technologies Inc.

Woodall, WH (2006). 'The use of control charts in health-care and public-health surveillance (with discussion).' *Journal of Quality Technology*, 38(22): 89–134.

3.3 *Attribution and causality in health-care performance measurement*

DARCEY D. TERRIS, DAVID C. ARON

Introduction

> *The important issue is that a good-quality indicator should define care that is attributable and within the control of the person who is delivering the care.*
>
> (Marshall et al. 2002)

A desirable health-care performance measure is one that reliably and accurately reflects the quality of care provided by individuals, teams and organizations (Pringle et al. 2002). The means of attributing causality for observed outcomes, or responsibility for departures from accepted standards of care, is critical for continuous improvement in service delivery. When quality measures do not reflect the quality of care provided then accountability for deficiencies is directed unfairly and improvement interventions are targeted inappropriately. It is both unethical and counterproductive to penalize individuals, teams or organizations for outcomes or processes outside their control.

In addressing attribution in health-care performance measurement, assessors must first face their own imperfections – specifically the likelihood that fundamental attribution error may influence quality assessments. Identified through social psychology research, fundamental attribution error occurs as a result of inherent human bias that arises when viewing another person's actions (Kelley 1967; Ross 1977). Specifically, causality is attributed to their behaviour by over-emphasizing an individual's disposition and under-emphasizing situational factors. This bias reflects a widespread cultural norm focusing on individual responsibility and free will that is reinforced by some legal frameworks.

When medical errors occur, it may be easier to recognize the active error that transpires rather than the multiple system-level errors that underlie it (Reason 2000). These latent errors may be more subtle and therefore more difficult to uncover and understand, especially in complex health-care environments. Even when latent errors are exposed, fundamental attribution error can lead us to ignore them and focus blame on the active error. This is problematic as failure to address the latent errors may provide fertile ground for future active errors. Given the tendency for fundamental attribution error, it is critical that health-care performance measurement is designed with scientific rigour. This is especially true when performance measures are linked to consequences (e.g. in reputation or reimbursement) that influence future service delivery. Perceived or experienced fundamental attribution error may lead to unintentional reductions in future health-care quality and equity (Terris & Litaker 2008).

For the purposes of performance measurement, a health outcome is said to be attributable to an intervention if the intervention has been shown in a rigorous scientific way to cause an observed change in health status. The mechanisms and pathways by which the intervention produces the change may not be known but there is some degree of certainty that it does. In this way much understanding of the world derives from experience-based causality, with statistical analysis providing support for the conclusions.

When attributing causality to a given factor or series of factors, typically a change in outcome is observed from manipulating one factor and holding all other factors constant. *Ceteris paribus* thus underlies the process and is a key principle for establishing models of causality. However, a strict *ceteris paribus* approach often cannot be obtained in the real world of health care. For example, when attributing clinical results in chronic disease management many factors outside the physician's actions are potentially involved. The interaction of these many factors (Fig. 3.3.1) further complicates the analysis. Definitive clinical outcomes may take years to manifest or occur so infrequently as to require large sample sizes to ensure detection with any degree of precision. Finally, random variations and systematic influences must be taken into account when differences in measured performance are being interpreted.

This chapter describes the challenges associated with assessing causality and attribution in health-care performance measurement and

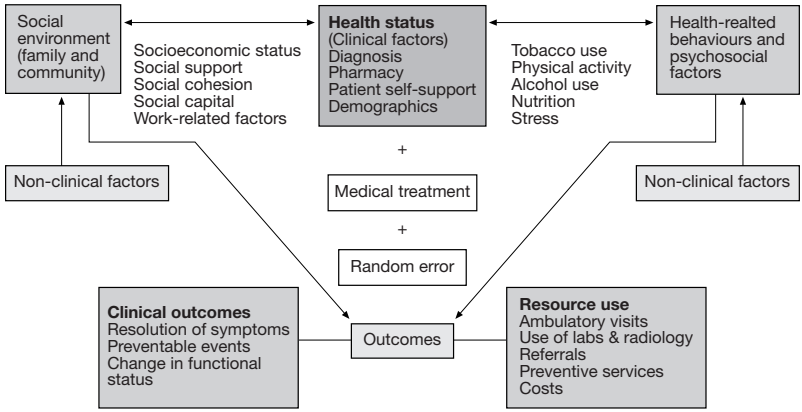

Fig. 3.3.1 Interrelationships of risk factors: relating risks to outcomes*

* Diagnosis-based measures are based on diagnoses, demographics and resource-use outcomes. Patient self-reported approaches are based on patient self-reported information (eg. health-related quality of life) and clinical outcomes.

The model shows that many factors outside a physician's actions can potentially influence the obtainment of a desired outcome of care. The number and interaction of these many factors complicates health-care performance measurement.

Source: Rosen et al. 2003

suggests methods for achieving at least a semblance of holding everything else constant. The concepts within the chapter are offered within the framework of performance measurement of health-care providers but are applicable to quality assessment at other levels including multi-provider practices, health-care facilities, hospitals and health systems. It is important to recognize that the methods presented rest upon a number of key assumptions. Specifically, most of our discussion is based on an underlying assumption of linear causality in which model inputs are assumed to be proportional to outputs. A critique of this approach is provided at the end of the chapter.

Assumptions underlying performance measurement

Donabedian's (1966) classic work on quality assessment identifies three types of performance measures – outcome, process and structure. Of these, outcome and process measures are most commonly used in health-care quality assessment. The reliability and accuracy of performance measurement requires proper definition (operationalization) of

the outcome and/or process under evaluation and the availability of good quality data. These are often the first assumptions made and it is dangerous to presume that either or both of these requirements are met.

It is assumed that the outcome or process under evaluation depends upon a number of factors. Iezzoni (2003) uses the phrase 'algebra of effectiveness' to describe health-care outcomes as a function of clinical and other patient attributes, treatment effectiveness, quality of care and random events or chance.

Patient outcomes = f (effectiveness of care or therapeutic intervention, quality of care, patient attributes or risk factors affecting response to care, random chance)

Each of these domains can be parsed in a variety of ways. For example, patient attributes may include clinical and health status parameters; health behaviours; psychosocial and socioeconomic factors; and individual preferences and attitudes. Effectiveness of care relates to the likelihood that a given intervention will result in the desired outcome e.g. that glycaemic control in a diabetic patient will reduce the occurrence of end-organ complications. Quality of care includes everything attributable to the delivery of health care whether at the physician, nurse, team or organizational level. This includes both the actions of the health-care providers and the context in which they practice. Finally, there are the vagaries of chance – the 'correct' therapy may not work for all patients.

Reliable and accurate assessment of a provider's role in health-care quality is dependent on the ability to divide and assign fairly the responsibility for a patient's receipt of appropriate services and attainment of desired outcomes to the many factors with potential influence. First, it must be known that a provider's given action or inaction *can cause* a process or outcome of care to occur. Then it must be ascertained whether (under the given circumstances and context) an observed process or outcome of care *is attributable* to the provider. The requirement for both causality and attribution implies that a provider's action/inaction may be neither 'necessary' (required to occur) nor 'sufficient' (needs presence of no additional factors in order to

occur) for a given process or outcome of care to transpire. Other factors, alone or in combination with the provider's action/inaction, may also cause the observed process or outcome of care to take place.

Similar issues may arise when using process measures even though receipt of a specific guideline recommended therapy (for example) would seem likely to avoid these uncertainties. A patient might not receive a guideline recommended therapy if the provider neglects to prescribe it. Conversely, the observed lack of therapy may occur if a provider prescribes the treatment but the patient refuses treatment because of his/her health beliefs. As illustrated, the provider's failure to prescribe is not 'necessary', i.e. the only possible cause for the observed absence of recommended therapy.

The level of attribution is also important. The provision of guideline-specified screening may occur as a result of a provider's knowledge and attention to standards of care. However, an automatic reminder system in the electronic medical record system utilized by the provider's practice may support the provider's memory and contribute to the observed rate of screening. In this case, the provider's memory alone is not 'sufficient'.

If a provider's actions/inactions are often neither necessary nor sufficient to cause an observed process or outcome of care, how is it possible to assess when the observed process or outcome of care can be ascribed, at least in part, to the provider? Statistical modelling through regression analysis is typically used to evaluate whether a significant relationship exists between providers and a process or outcome variable identified as a quality indicator. Through a process of risk adjustment, control variables are included in the model to account for the potential effects of other factors (confounders) that may influence the incidence of the quality indicator under investigation.

However, even with risk adjustment, more than a single model is necessary to prove that an observed quality indicator is causally linked and attributable to a provider's action/inaction. Measurement and attribution error, complexity in the confounding relationships and provider locus of control must be considered in the analysis of causality and attribution for health-care performance measures (Fig. 3.3.2). The risks associated with causality and attribution bias and the methods to reduce such bias are explored in this chapter.

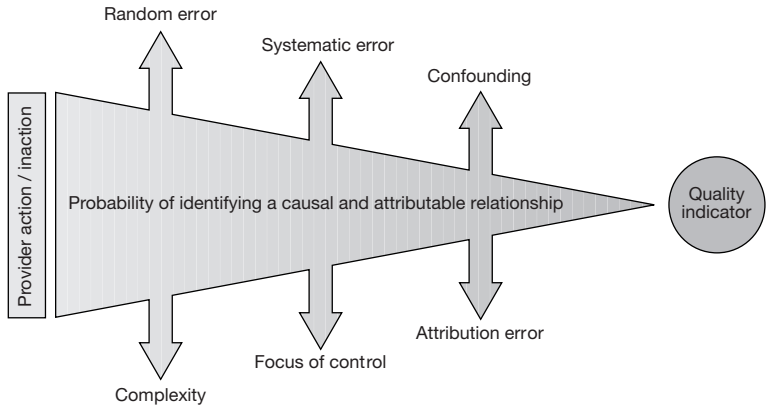

Fig. 3.3.2 Health-care performance measurement: challenges of investigating whether there is a causal and attributable relationship between a provider's action/inaction and a given quality indicator

The vagaries of chance in health-care performance measurement – random error

Variability arising from chance or random error is present in all quantitative data. Two types of random error must be considered in statistical estimates, including those employed in health-care performance measurement. The first is commonly referred to as type I error, or the false positive rate; the second is called type II error, or the false negative rate. Individual variables may be subject to higher or lower rates of random error. For each variable, the errors happen at random without a systematic pattern of incidence within the data elements collected. However, the variance falls evenly above and below the true value of the variable being measured. With increasing random error, the mean value for the variable is unaffected although the variance will increase. In general, variance decreases with increasing sample size.

The acceptable type I error rate of a statistical test (also called the significance level or p value) is typically set at 0.05 or 0.01. This is interpreted to mean that there is a five in one hundred or a one in one hundred chance that the statistical test will indicate that a relationship exists between two variables under consideration (e.g. a provider's action/inaction and a quality indicator) when a relationship *is not present*. Therefore, even when the results of statistical modelling

suggest a significant relationship between two variables, it must be recognized that there is a chance that the conclusion is false.

Further, with repetitive testing there is an increasing likelihood that type I error will produce one or more false conclusions unless the analyses adjust for this risk (Seneta & Chen 2005). This problem is especially prevalent in quality measurement due to the proliferation of individual measures and multiple comparisons. Under these circumstances, it may be more common than is acknowledged to see a significant relationship that truly does not exist (Hofer & Hayward 1995 & 1996).

Researchers may also fail to detect differences that are present, i.e. a false negative result may occur. In general, there is more willingness to accept a false negative conclusion (type II error) than a false positive conclusion (type I error). Therefore, the type II error rate (ß) is typically set in the range of 0.20 or 0.10. With ß = 0.20, there is a 20% chance of a conclusion that there is no relationship between two variables when a relationship *does* exist. Statistical testing does not usually refer directly to the type II error rate and the power of the test (1- ß) is more commonly reported. Power analysis is performed before data are collected in order to identify the size of the sample required. This increases the likelihood that the desired type II error rate will not be exceeded. When performed after data collection and statistical testing, power analysis identifies the type II error rate achieved. If the type II error rate is greater than the desired rate, a study may be described as under-powered.

It is not possible to reduce the risk of type I and II error simultaneously without increasing sample size. Sample size may be increased by merging data from smaller units or across time, or through a combination of these approaches. Increasing sample size by these methods may reduce the impact of chance but may also change the focus of the analysis. The results from the aggregated data may be less useful for assessing the health system level and/or time period of interest.

A pervasive statistical phenomenon called regression to the mean may also make natural variation in repeated data look like real change (Barnett et al 2005; Morton & Torgerson 2005). When data regress to the mean, unusually high (or low) measurements tend to be followed by measurements that are closer to the mean. Statistical methods can assess for regression to the mean but have not been used to any great extent (Hayes 1988).

Greater variance from chance (random error) in data makes it more difficult to draw a conclusion as to whether a relationship exists between two variables under analysis. All data are subject to random error which can be minimized through careful adherence to measurement and data recording protocols; with routine checks of data reliability and completeness; and through the use of control groups when possible.

Systematic error in health-care performance measurement

The certainty associated with an estimate of the relationship between two variables is also subject to systematic error. This is also called inaccuracy or bias and results from limitations in measurement and sampling procedures. Systematic error may occur when all measured values for a given variable deviate positively or negatively from the variable's true value, for example – through poor calibration of the measurement instruments employed. This type of bias would equally affect all members of the sample, resulting in the mean for the sample deviating positively or negatively from the true population mean. Bias may also occur when erroneously higher (or lower) values for a given variable are more likely to be measured for a subgroup under analysis. This can occur in resource-limited settings where the measurement instruments used by providers are more likely to be out of calibration than those used in resource-affluent settings.

As with random error, there is no way to avoid all sources of systematic error when assessing the presence of a relationship between two variables. Unlike random error, however, it is not possible to set a maximum rate of permitted systematic error when drawing statistical conclusions. Assessments of systematic error are not included routinely in reports of statistical results (Terris et al. 2007) but recently there has been greater attention to the need for routine, quantitative estimation of bias and its effect on conclusions drawn in statistical analyses (Greenland 1996; Lash & Fink 2003; Schneeweiss & Avorn 2005).

Systematic error obscures assessment of the size and nature of the relationship between two variables. For example, the presence of bias may lead to the conclusion that the relationship between a provider's action/inaction and a given quality indicator is larger (or smaller) than the actual association. Under these circumstances, more (or less) operational significance may be assigned to the identified relationship.

Systematic error can be reduced by proactively considering potential sources of bias in the design and implementation of measurement systems. This enables protocols to be implemented to minimize systematic error in measured values and limit bias among study subgroups.

Confounding in health-care performance measurement

If careful data collection and statistical tests have produced confidence that a relationship exists between two variables under consideration, is it then possible to assume that the relationship is causal? Unfortunately, a significant statistical result only implies that a causal link may be present – it does not prove causality and the relationship can only be said to be correlative. Correlated variables move together, or co-vary, in a pattern that relates to each other. Positive correlation exists when the variables move together in the same direction; negative correlation exists when the variables move in opposition to each other. In both instances, the underlying drivers of the association between the two variables remain unknown.

Correlated variables may be causally linked to each other or both variables under consideration may be affected by a third variable, called a confounder. When the relationship between two variables is confounded by a third variable, the third variable may cause all or a portion of the observed effect between the first two. The confounder's common influence on the first two variables creates the appearance that these two are more strongly connected than they are.

Multivariate statistical modelling controls for confounding by including factors with potential influence on the observed relationship between the primary hypothesized causal agent and the process or outcome variable of interest. This process of controlling is called risk adjustment. The identification of possible confounders and specification of models to control adequately for their effect in health-care performance measurement is discussed in detail in Chapter 3.1.

If an analysis does not adequately account for confounding then the estimated relationship between the two variables of interest will be biased. This type of bias is called missing variable or misspecification bias. As discussed, bias in an assessment of the relationship between two variables can lead to the conclusion that the relationship between the two variables is larger (or smaller) than the actual association. A positive relationship might even be construed as negative, or vice versa.

Complexity in health-care performance measurement

Within a given health-care delivery context, the complexity arising from the number of potential confounders and the complicated relationships between possible confounding factors creates a daunting challenge when seeking to attribute an observed process or outcome of care to a provider's action/inaction. However, variation due to other causes must be accounted for before an observed process or outcome of care can be attributed to a provider's action/inaction (Lilford et al. 2004). Possible confounders arise from patient-level characteristics as well as the health-care resources, systems and policies surrounding the patient and the patient-provider encounter (Rosen et al. 2003; Terris & Litaker 2008). This is further complicated by the need to consider potential confounders that arise outside the health-care environment (see Box 3.3.1 for an example). Adequate risk adjustment for potential confounders is limited by both the knowledge and acknowledgement of potential confounding agents and the ability and available resources to capture confounders for inclusion in quality assessments.

Box 3.3.1 Community characteristics and health outcomes

Empirical studies suggest that community and neighbourhood-level factors have an impact on the health status and outcomes of residents. These factors include the neighbourhood's socioeconomic status; physical environment and availability of resources (recreational space, outlets to purchase fresh foods, etc.); and the social capital within the community. These effects are linked to the context in which people live, not the people themselves (Litaker & Tomolo 2007; Lochner et al. 2003).

For example, Lochner et al. (2003) used a hierarchical modelling approach to demonstrate that neighbourhoods with higher levels of social capital (as assessed by measures of reciprocity, trust and civic participation) were associated with lower all-cause and cardiovascular mortality. This result was found after adjusting for the material deprivation of neighbourhoods. Therefore, individuals living in neighbourhoods with lower social capital may be at greater risk of poor health outcomes, regardless of the quality of care given by their providers.

This discussion can be extended by returning to the previous example in which a patient does not receive a guideline-specified treatment. If the receipt of treatment is used as a quality indicator, this episode reflects negatively on the provider and will be classified as an instance of poor quality care. However, as previously discussed, the patient's health beliefs may have led him/her to refuse the prescribed treatment. Conversely, the patient may have been willing to follow the recommendation but access to the therapy was restricted by policies set by their health-care coverage agency. Limitations in the availability and capacity of facilities dispensing the treatment may also have created insurmountable barriers for the patient. Finally, the patient could have received the treatment but this was not recorded in the health information systems in place (see Box 3.3.2 for a further example). These are just a few of the many factors that may have influenced the observed failure to receive the guideline-recommended treatment, outside of the provider's failure to recommend the therapy.

As the hypothetical example shows, confounding factors that influence an observed process or outcome of care can originate from

Box 3.3.2 Missed opportunities with electronic health records

By reducing barriers to longitudinal health and health-care utilization information, electronic health records (EHRs) can be used to improve the quality of care delivered to patients and the reliability and validity of health-care performance measurement. However, in a recent study by Simon et al. (2008) less than 20% of the provider practices surveyed (in Massachusetts, USA) reported having EHRs. Of those practices without, more than half (52%) reported no plans to implement an EHR system in the foreseeable future. Funding was the most frequently reported obstacle to implementation.

Further, less than half of the systems in practices with EHR systems provided laboratory (44%) or radiology (40%) order entry (Simon et al. 2008). This misses the opportunity to, for example, identify whether a provider ordered a guideline-recommended laboratory test. The only information available to assess the quality of care delivered would be the absence of the test result. If the patient did not receive the test for reasons outside the provider's control, this scenario would reflect unfairly upon the provider.

several levels within the health-care delivery environment. In the example given, the confounder was hypothesized to have arisen from patient-level characteristics (patient's health beliefs); provider practice resources (information systems); health system policies (reimbursement policy); or the patient's home community (capability and accessibility of dispensing facilities).

In health-care performance assessment, providers can be sorted into subgroups at different levels, for instance – based on the facilities they practice within; the coverage programmes in which they are included; and/or the communities they serve. The actions/inactions of providers within a given subgroup (e.g. providers practising at a given hospital) tend to have less variation than the actions/inactions of providers in different subgroups (e.g. providers practising at separate hospitals). Hierarchical models can be used to differentiate between the variation arising from differences between providers and between subgroups of providers. If the clustering of data is not accounted for then the estimate of the relationship between the provider's action/inaction and the quality indicator may be biased. Further, the confidence intervals (i.e. estimated range of the effect of the providers' action/inaction on the quality indicator, based on the significance level of the test) may also be narrowed, leading to false conclusions concerning the apparent significance of the relationship (Zyzanski et al. 2004). Therefore, hierarchical modelling approaches have been increasingly recommended (Glance et al. 2003).

Provider locus of control

The example discussed above raises the issue of access hurdles that may prevent a patient from following a provider's recommended therapy. From the provider's perspective, these same hurdles may functionally limit their own control of care-delivery recommendations. For example, health system policies may restrict the number of referrals that a provider can make within a given period. Non-emergency patients who present at the provider's office after the referral limit has been reached may be requested to return for a referral at a later date. However, performance assessment for the time of the postponement would indicate that the recommended process of care had not occurred.

Health system policies may also encourage providers to pursue therapies other than their preferred course of treatment. The new

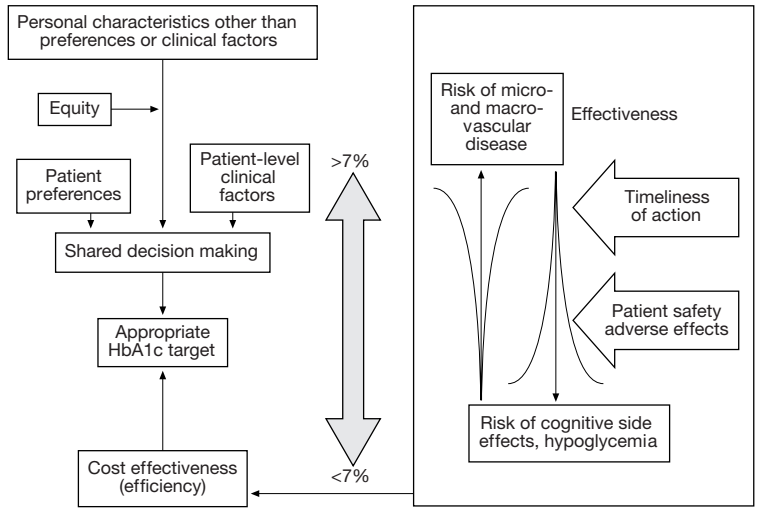

Fig. 3.3.3 Factors in choice of target HbA1c for care of a given patient with diabetes

Source: Based on the model by Aron & Pogach 2007

diabetes care quality measure adopted by the National Committee for Quality Assurance (NCQA) can be used to illustrate this point. The measure is based on the percentage of diabetic patients aged eighteen to seventy-five who have HbA1c levels of less than 7% (Pogach et al. 2007). This target HbA1c level may indicate excellent glycaemic control but a number of factors should be considered before choosing a target HbA1c for a given patient. A conceptual framework illustrating these factors is shown in Fig. 3.3.3.

For example, consider a seventy-four-year-old man with diabetes and heart failure who takes oral medications for glycaemic control. He would require insulin injections to improve his glycaemic control from an HbA1c of 7.2% to less than 7%. However, these injections would increase the patient's risk of hypoglycaemia and its attendant morbidity with little benefit in terms of reduction in cardiovascular risk or microvascular complications. Further, the patient may strongly prefer to continue with the oral agents. Should this patient be counted against his provider because the HbA1c quality target is not met? Of more concern, should the health system's policies lead the provider

to strongly recommend (coerce?) the patient to accept insulin injections in order to meet the quality target?

A provider's locus of control can be significantly affected by the policies and infrastructure of their practice environment as directed by local, regional and national health systems and regulatory bodies (Hauk et al. 2003; Landon et al. 2001). Even if a causal relationship is established between a provider's action/inaction and a performance indicator, the responsibility for an observed process or outcome of care may not always be attributable to the provider. Further, providers' locus of control may vary substantially between different practice contexts and for different patient subgroups within a given context. Factors that influence a provider's ability to direct their actions/inactions within their practice environment should be accounted for in health-care performance measurement. These factors are possible confounders to be included in the risk adjustment process.

Attribution theory and fundamental attribution error

Much has been said about the complexity encountered when trying to establish a causal link and attribute a provider's action/inaction to an observed care process or health outcome. It may be that health-care quality researchers have over-emphasized this complexity due to fundamental attribution error. Originating in social psychology research, the term is used to describe bias that arises from differences in perspective when identifying the causal factors for events in which we have been involved and events concerning others (Jones & Harris 1967; Ross 1977). Specifically, there is a known tendency to over-emphasize situational factors (those outside ourselves) when looking for explanations of outcomes related to our own actions. Conversely, when looking at others we are predisposed to under-emphasize these same situational factors and focus more on individual responsibility.

For example, in a recent study by Golomb et al. (2007), physicians were reluctant to attribute patient-reported symptoms to an adverse effect of drugs that they had prescribed. This hesitation occurred even when the reported symptom had strong literature-based support for probable drug causality. Within the framework of fundamental attribution error the physicians could be unconsciously reluctant to attribute reported symptoms to their decision to prescribe the drug. Further, they may be more likely to attribute the reported symptoms

to the patient's other health behaviours, downplaying the influence of the external factor of the drug's potential side effects.

Similarly, it might be hypothesized that insiders involved in developing performance measurement systems in a health-care system are more apt to look for external factors as possible confounders. Conversely, when outsiders investigate these performance measurement systems they may be less likely to include external factors as possible confounders. The outsiders may focus instead on the personal responsibility of the agent (e.g. providers, hospitals) under analysis. To limit the effect of fundamental attribution error on the development of health-care performance measures, causality and attribution should be assessed with scientific rigour. Multiple perspectives should be included in the analysis by involving internal and external stakeholders.

Causality and attribution bias in health-care performance measurement

When there is bias in the assessment of causality or attribution for a given quality indicator for a subgroup of providers, the affected providers are ranked more advantageously or disadvantageously (with respect to their true quality) than providers of corresponding quality. If reimbursement is linked to health-care performance assessment then providers subject to this bias are unfairly compensated, receiving a higher or lower rate of payment than providers of equivalent quality. If market-share incentives are offered through published public scorecards, providers who have experienced bias in their assessment will appear relatively more or less attractive to patients than providers of similar quality (Lilford et al. 2004).

Both providers and patients bear the risk of causality and attribution bias in health-care performance measurement. Providers are treated unfairly: well-compensated regardless of the relatively poor quality of care provided or penalized despite the relatively high-quality service delivery. As a consequence, patients may receive lower quality health care. They may leave a relatively high-quality provider because public reporting has misrepresented the provider as delivering low-quality care or because the provider has instituted restrictions in their practice in response to lower reimbursement rates based on this inaccurate assessment.

Who is at risk from causality and attribution bias?

Providers who practise in resource-limited settings are at greater risk of bias in health-care performance measurement than their counterparts in more resource-affluent settings (Casalino & Elster 2007; Terris & Litaker 2008). This bias arises, in part, from differences in the provider's locus of control in acquiring and directing the use of resources in the delivery of care. When resources are generally limited, the choices available to the provider are also limited. The resources to be considered include those that providers apply in service delivery, specifically the facilities, equipment, personnel, management and information systems available (Miller & West 2007).

Limitations in community resources (e.g. neighbourhood's socio-economic status; local public health policy and practice; general infrastructure) may also increase the risk that a provider practising within the community will be subject to bias in health-care performance measurement. These community-level factors influence the health and health-care processes and outcomes obtained by residents. Providers that service resource-limited settings also typically face greater complexity in their practice and this may be difficult to capture and include when risk adjusting in the health-care assessment process (Casalino & Elster 2007). Sources of information outside the practice (such as community-level economic data) are necessary to account adequately for the complexity of the practice context.

Providers that care for more complex patients are also at greater risk of bias in health-care performance measurement (Terris & Litaker 2008). This complexity can arise from the health status of the patient (e.g. severity; comorbidity) or from other patient-level characteristics (e.g. socio-economic status; health beliefs and behaviour). Providers that practise in resource-limited settings generally treat a greater proportion of complex patients (Casalino & Elster 2007). However, complex patients are also more likely to be found within the practices of providers affiliated to teaching hospitals (Antioch et al. 2007) or who specialize in more complex patient subgroups, such as frail older adults (Jette et al. 1996). Risk adjustment for severity and comorbidity is common but again other sources of information are necessary to incorporate the additional patient-level factors that can influence the obtainment of desired processes and outcomes of care.

It should be noted that the bias in health-care performance measurement that arises from limitations in sources and the quality of data and reporting systems is not restricted to providers in resource-limited settings (Terris & Litaker 2008). First, regardless of general resources, few providers have access to or utilize more technologically advanced information systems such as electronic health records (Burt & Sisk 2005). Second, more sophisticated information systems for data recording and reporting are not guaranteed to capture reliably and accurately the patient-level factors that are accessible within the patient-provider encounter (Persell et al. 2006). For example, an electronic health record might not have a clear entry point for specific information on a patient's less common contraindication for a guideline recommended treatment (e.g. patient states he/she is unable to swallow pills). However, a written medical record can afford the provider greater flexibility to note this confounding factor.

What are the potential effects of causality and attribution bias on health-care quality and equity?

Performance measurement is used by health-care managers to both identify targets for improvement and incentivize providers to improve service delivery (Terris & Litaker 2008). If the causality and attribution of a provider's action/inaction to a given quality indicator is not assessed accurately and reliably then the signal that this action/inaction should be repeated (or avoided) will be lost. A high-quality provider may not sustain their current practice policies and procedures as they would not link their current routines with the delivery of high-quality care. As a consequence, new initiatives may be substituted that may/may not result in a similar or better level of care. Conversely, a relatively low-quality provider that is assessed inaccurately as providing higher-quality care does not receive the clear signal that service delivery needs to be improved. The opportunity to maintain and improve quality is clearly affected when providers experience bias in health-care performance measurement.

When health-care performance measurement is linked to reimbursement or other market-based incentives, providers' perception of the risk associated with inaccurate assessment may create disincentives that are contrary to the goal of improving equity in access and health-care quality (Lilford et al. 2004). Providers may seek to avoid

including complex patients in their practice or locating their practice in more complex settings. This disincentive may create difficulties in the recruitment and retention of providers for disadvantaged population segments. Further, providers that deliver high-quality care in resource-limited settings may be reimbursed at a lower rate than those that supply a similar level of care in a more resource-affluent environment. This would lead to further restrictions in the resources available for health care in resource-limited settings and likely degradation in the quality of care delivered. In this manner, biased health-care performance measurement could result in increased health disparities (Casalino & Elster 2007).

The probability of fundamental attribution error increases with the increasing risk linked to health-care performance measurement. A provider with a reputation, reimbursement rate or market share at stake may be more likely to point to factors outside their locus of control as responsible for the observed process or outcome of care. Future opportunities for quality improvement are lost as the fear of penalties (fair or unfair) leads providers to avoid self-reflection and instead to identify external causal agents.

Methods to reduce causality and attribution bias in health-care performance measurement

The one certainty in health-care performance measurement is that most often it will not be known with absolute certainty that an observed process or outcome of care is causally linked and attributable to a provider's action/inaction. However, it is possible to address actively the risk of bias in the assessment of causality and attribution in the development and implementation of measurement systems in order to maximize the certainty obtained. A first step is proactive consideration of the possible pathways leading to the desired process and/or outcome of care and where they can diverge (Institute of Medicine 2007).

In industry, failure mode and effects analysis and root cause analysis are advocated during the product or process development stage in order to anticipate risks for adverse events and the need for process control points (McCain 2006). In health care, these methods are increasingly applied to improve patient safety but are most commonly retrospective, in response to an adverse event or near miss (Battles et al. 2006). For example, root cause analysis has been systematically applied for

adverse events and near misses that occur in Department of Veterans Affairs' medical facilities. Implementation of this process has shifted the focus from human errors to system vulnerabilities and more actionable root causes (Bagian et al. 2002). Proactive examination of the pathways to episodes of high- and low-quality care enables even more comprehensive understanding of the provider's role and identification of possible confounders, their potential impact and the probability of their influence within a given context. Research can then be designed to investigate whether there is a causal and attributable relationship between a provider's action/inaction and a given process or outcome of care.

The randomized controlled trial (RCT) is considered the gold standard in study design for clinical evidence and has been advocated for building the necessary evidence for quality of care research (Institute of Medicine 2007). Study subjects are assigned randomly to either a treatment (e.g. provider's action) or control (e.g. provider's inaction) group. If the study sample is sufficiently large, random assignment will result in an equal distribution of possible confounders between the treatment and control groups. However, random assignment can account only for confounders represented among the subjects in the study sample. The representativeness of the study sample to more general populations and alternative health-care delivery contexts must be assessed before extending the results of the RCT. Random assignment to a treatment or control group may be neither possible nor ethical in all study scenarios. This may be especially true in quality of care research in which it may be known that a given provider action is preferable (i.e. the action does no harm and may result in benefit) but not whether the action is causal or attributable in a given health-care delivery context. Under such circumstances it may be considered unethical to withhold a potentially beneficial action from study participants (Edwards et al. 1998).

Well-designed observational study designs can be used when an RCT is not possible. Observational studies are potentially affected by hidden bias and therefore sensitivity analyses should be performed routinely in the assessment of results. Propensity score (Johnson et al. 2006) and instrumental variable (Harless & Mark 2006) methods are also recommended increasingly in the analysis of observational study results (see Box 3.3.3 for examples). Propensity score and instrumental variable methods are used to approximate the randomization process of an RCT.

Box 3.3.3 New views on the volume-outcome relationship

Numerous studies have identified a link between the volume of health care delivered and patient outcomes, with higher volume hospitals and providers appearing to provide higher quality of care (Halm et al. 2002). However, prior analysis of the volume-outcome relationship may have been confounded in two important ways. First, the studies may not have risk adjusted adequately for differences in the case mix of patients attending high- and low-volume providers. Second, the relationship may actually be one of reverse causality (Luft 1980). Higher volume may not lead to better outcomes (practice-makes-perfect argument) but providers who are associated with better outcomes may receive more referrals.

New evidence using propensity scores to adjust for selection bias
Zacharias et al. (2005) used a propensity score approach to address systematic differences in patient characteristics before comparing CABG outcomes between a high- and a low-volume hospital. Propensity scores were derived from a logistic regression model, with presentation at the high- or low-volume hospital as the dependent variable. A wide variety of patient-level risk factors were included as covariates. The model was then used to calculate a propensity score for each patient included in the sample. Patients were matched (one from each hospital) based on their propensity score and their CABG outcomes were compared. In the final analysis, hospital volume was not found to be a significant predictor of in-hospital mortality or three-year survival.

Further evidence using an instrumental variable approach
Tsai et al. (2006) used an instrumental variable approach to investigate the volume-outcome relationship among inpatients with congestive heart failure. The instrumental variable used was the linear distance between a patient's residence and the hospital in which care was received. This distance is conceivably related to the exposure of interest (hospital volume, with patients more likely to attend closer hospitals) but not the outcome of interest (thirty-day mortality). The researchers repeated their analysis using limited administrative data and more complete clinical data for risk adjustment and

Box 3.3.3 cont'd

including and not including the instrumental variable in the model. A small, potential volume-outcome relationship was only found when the limited administrative data were used and the instrumental variable was not included in the final model. A significant relationship was not found under the other model scenarios.

Propensity scores are derived through multivariate logistic regression models, using receipt of the exposure (e.g. provider's action/inaction) as the outcome variable and factors that influence the receipt of the exposure (e.g. measures of the patients' health status) as covariates. The goal is to include all variables that play a role in receipt of the exposure in order to model propensity for exposure. The model should include interactions among identified covariates although it appears that it is more important to include all the relevant predictors than the correct interaction terms (Dehijia & Wahba 1998; Drake & Fisher 1995).

The exposure model is used to derive a propensity score for each patient, based on the patient's status for each covariate included. Next, patients who did and did not receive the exposure are matched according to their propensity scores. This approximates equal distribution of confounders associated with receipt of treatment between a treatment and control group in an RCT. In a second stage of regression analysis, differences in the outcomes observed between propensity score-matched subjects are then attributed more accurately to the exposure. Propensity score methods work best with large samples and where data are collected expressly for the purpose of deriving propensity scores for subject matching. They can adjust only for measured covariates associated with receipt of exposure and not for unmeasured or omitted variables (Braitman & Rosenbaum 2002). Therefore, the more intensive data collection required for these methods may not be suitable for routine quality assessment.

Instrumental variable models are recommended when there is potential feedback between the outcome (e.g. quality indicator) and exposure (e.g. a provider's action/inaction); unmeasured confounders in the analysis; and/or significant measurement error. A selected instrumental variable should be associated with the exposure vari-

able but not the outcome variable. When the instrumental variable is included in the regression analysis it will appear to be associated with the outcome variable because of its relationship with the exposure variable. The association identified between the instrumental and the outcome variable can then be divided into the association between (i) the instrumental and exposure variable; and, more importantly, (ii) the exposure and outcome variable.

Other techniques have also been developed to address complexity in the assessment of causality and attribution. These methods include multi-level modelling to separate out the hierarchical effects associated with clustered data (Leyland & Goldstein 2001) and selection bias models (Weiner et al. 1997). To date, none of these methods has been widely adopted in health-care performance measurement.

Beyond study design and statistical technique, it is important to recognize that a single RCT or well-designed observational study does not provide sufficient evidence of causality between a provider's action/inaction and a process or outcome of care. A preponderance of evidence is needed from multiple studies among different sample populations and service-delivery contexts. If a plausible pathway is hypothesized and supported through such research results then greater certainty can be assigned to the identified causal link. Further, this derives a richer picture of the health-care delivery contexts in which the process or outcome of care is attributable to the provider's action/inaction, leading to possible multi-factorial interventions to improve future quality.

Critique from the standpoint of complexity theory

The foundation of evidence-based medicine relies upon a particular conceptual model of the world. This model describes a mechanistic world that functions according to deterministic principles in which problems are analysed using a framework of simple linear causality. To illustrate this point, consider an environmental toxin associated with a particular cancer (e.g. aflatoxin and liver cancer). Under the assumption of linear causality, it is presumed that the effect (liver cancer) can be eliminated by eliminating the cause (exposure to aflatoxin). However, health effects are generally not caused by a single agent – there is a web of causal factors, of which the effect itself may be a

part. This view is grounded in complexity theory and the behaviour of complex systems.

Complex systems comprise a large number of interacting components that have interconnecting actions. They contain many direct and indirect feedback loops and so the interactions are non-linear with non-proportional effects. Small changes can have large effects on overall system behaviour while large changes can have little effect. The behaviour of the system is determined by the nature and effect of the interactions, not solely by the content or individual actions of component elements (Rouse 2000 & 2003).

If health systems are accepted as complex systems under this definition, there must be a fundamental revision of the understanding of causality and attribution as described within this chapter. Further, the methods used to identify targets and implement health-care quality improvement initiatives will change radically. Until that time, it will be necessary to rely upon the simpler models presented, focusing on individual causal agents but acknowledging the context and systems within which they work.

Conclusions

Health-care managers involved in health-care performance measurement are advised to consider the following recommendations in addressing causality and attribution bias.

1. Access existing reports of research into the possibility of a causal and attributable link between the agents under assessment (e.g. providers, hospitals) and the process or outcome of care proposed as a quality indicator. Evaluate the quality of this research based on study design and control for confounding. Context is important as findings based on a given patient population or setting (health-care venue or system; social, cultural or economic environment; etc.) may not be generalizable to other contexts or countries.
2. Perform a prospective analysis to identify the critical pathways involved in the achievement of desired and undesired processes and outcomes of care. Identify possible confounders to the relationship between the agents under assessment and the process or outcome of care proposed as a quality indicator. Further, identify

how the agents under assessment may be clustered within levels of the health-care context under analysis.

3. Synthesize the results of steps 1 and 2 and identify essential gaps in knowledge. Involve stakeholders internal and external to the health-care level under analysis in order to minimize the risk of fundamental attribution error. Consider root cause analysis as a method to identify system-level sources of variation in the quality of care delivered. These root causes may be more effective targets for sustainable improvement efforts.
4. If a new study is required, prospectively consider sources of random and systematic error in measurement and sampling when developing the study design. This applies to studies utilizing either primary or secondary data sources. Institute policies and procedures for data collection that maximize the reliability and accuracy of the data used for the quality assessment. In resource-constrained settings, it may be more useful to employ a limited number of quality indicators that can be measured in a repeatable and valid manner rather than overburdening reporting mechanisms with many indicators that are less reliable and accurate.
5. Employ risk adjustment when evaluating the relationship between the agents under assessment and the process or outcome of care proposed as a quality indicator. Consider the use of hierarchical models to account for the clustering of data within levels of the health-care context under analysis (see step 2). When confounding cannot be controlled for through randomization, further consider the use of propensity score or instrumental variable methods to approximate randomization.
6. Acknowledge that causality and attribution bias cannot be eliminated completely, even when utilizing best practices as described above. Consider the unintended impacts from experienced or perceived bias in quality assessment on the future improvement of health-care quality and equity, especially when reimbursement or market-share incentives are linked to quality assessment. The risk and potential consequences of causality and attribution bias may be especially severe in resource-constrained and complex settings or for those who care for patients with more complex needs.

References

Antioch, KM. Ellis, RP. Gillett, S. Borovnicar, D. Marshall, RP (2007). 'Risk-adjustment policy options for casemix funding: international lessons in financing reform.' *European Journal of Health Economics,* 8(3): 195–212.

Aron, DC. Pogach, LM (2007). 'One size does not fit all: a continuous measure for glycemic control in diabetes: the need for a new approach to assessing glycemic control.' *Joint Commission Journal on Quality Improvement,* 33: 636–643.

Bagian, JP. Gosbee, J. Lee, CZ. Williams, L. McKnight, SD. Mannos, DM (2002). 'The Veterans Affairs root cause analysis system in action.' *Joint Commission Journal on Quality Improvement,* 28(10): 531–545.

Barnett, AG. van der Pols, JC. Dobson, AJ (2005). 'Regression to the mean: what it is and how to deal with it.' *International Journal of Epidemiology,* 34(1): 215–220.

Battles, JB. Dixon, NM. Borotkanics, RJ. Rabin-Fastmen, B. Kaplan, HS (2006). 'Sensemaking of patient safety risks and hazards.' *Health Services Research,* 41(4 Pt 2): 1555–1575.

Braitman, LE. Rosenbaum, PR (2002). 'Rare outcomes, common treatments: analytic strategies using propensity scores.' *Annals of Internal Medicine,* 137(8): 693–695.

Burt, C. Sisk, J (2005). 'Which physicians and practices are using electronic medical records?' *Health Affairs,* 24(5): 1334–1343.

Casalino, L. Elster, A (2007). 'Will pay-for-performance and quality reporting affect health disparities?' *Health Affairs,* 26(3): 405–414.

Dehejia, R. Wahba, S (1998). *Propensity score-matching methods for non-experimental causal studies*. Cambridge, MA: National Bureau of Economic Research (NBER Working Paper No. 6829).

Donabedian, A (1966). 'Evaluating the quality of medical care.' *Milbank Memorial Fund Quarterly,* 44(3 Pt. 2): 166–203.

Drake, C. Fisher L (1995). 'Prognostic models and the propensity score.' *International Journal of Epidemiology,* 24(1): 183–187.

Edwards, SJL. Lilford, RJ. Hewison, J (1998). 'The ethics of randomised controlled trials from the perspectives of patients, the public and healthcare professionals.' *British Medical Journal,* 317(7167): 1209–1212.

Glance, LG. Dick, AW. Osler, TM. Mukamel, D (2003). 'Using hierarchical modeling to measure ICU quality.' *Intensive Care Medicine,* 29(12): 2223–2229.

Golomb, BA. McGraw, JJ. Evans, MA. Dimsdale, JE (2007). 'Physician response to patient reports of adverse drug effects: implications for patient-targeted adverse effect surveillance.' *Drug Safety,* 30(8): 669–675.

Greenland, S (1996). 'Basic methods for sensitivity analysis of biases.' *International Journal of Epidemiology,* 25(6): 1107–1116.

Halm, EA. Lee, C. Chassin, MR (2002). 'Is volume related to outcome in health care? A systematic review and methodologic critique of the literature.' *Annals of Internal Medicine,* 137(6): 511–520.

Harless, DW. Mark, BA (2006). 'Addressing measurement error bias in nurse staffing research. *Health Services Research,* 41(5): 2006–2024.

Hauk, K. Rice, N. Smith, P (2003). 'The influence of health care organisations on health system performance.' *Journal of Health System Research & Policy,* 8(2): 68–74.

Hayes, RJ (1988). 'Methods for assessing whether change depends on initial value.' *Statistics in Medicine,* 7(9): 915–927.

Hofer, TP, Hayward, RA (1995). 'Can early re-admission rates accurately detect poor-quality hospitals?' *Medical Care,* 33(3): 234–245.

Hofer, TP. Hayward, RA (1996). 'Identifying poor-quality hospitals: can hospital mortality rates detect quality problems for medical diagnoses?' *Medical Care,* 34(8): 737–753.

Iezzoni, LI (2003). *Risk adjustment for measuring health care outcomes, Vol. 3*. Chicago: Health Administration Press.

Institute of Medicine (2007). State of the science of quality improvement research. In: *The state of quality improvement and implementation research: expert views. Workshop summary.* Washington DC: The National Academies Press.

Jette, AM. Smith, KW. McDermott, SM (1996). 'Quality of Medicare-reimbursed home health care.' *Gerontologist,* 36(4): 492–501.

Johnson, ML. Bush, RL. Collins, TC. Lin, PH. Liles, DR. Henderson, WG. Khuri, SR. Petersen, LA (2006). 'Propensity score analysis in observational studies: outcomes after abdominal aortic aneurysm repair.' *American Journal of Surgery,* 192(3): 336–343.

Jones, EE. Harris, VA (1967). 'The attribution of attitudes.' *Journal of Experimental Social Psychology,* 3: 1–24.

Kelley, HH (1967). Attribution theory in social psychology. In: Levine, D (ed.). *Nebraska symposium on motivation and emotion, Vol. 15*. Lincoln: University of Nebraska Press.

Landon, BE. Reschovsky, J. Reed, M. Blumenthal, D (2001). 'Personal, organizational, and market level influences on physicians' practice patterns: results of a national survey of primary care physicians.' *Medical Care,* 39(8): 889–905.

Lash, TL. Fink, AK (2003). 'Semi-automated sensitivity analysis to assess systematic errors in observational data.' *Epidemiology,* 14(4): 451–458.

Leyland, AH. Goldstein, H (2001). *Multilevel modeling of health statistics*. Chichester: John Wiley & Sons.

Lilford, R. Mohammed, MA. Spiegelhalter, D. Thomson, R (2004). 'Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma.' *Lancet,* 363(9424): 1147–1154.

Litaker, D. Tomolo, A (2007). 'Association of contextual factors and breast cancer screening: finding new targets to promote early detection.' *Journal of Women's Health,* 16(1): 36–45.

Lochner, KA. Kawachi, I. Brennan, RT. Buka, SL (2003). 'Social capital and neighborhood mortality rates in Chicago.' *Social Science & Medicine,* 56(8): 1797–1805.

Luft, HS (1980). 'The relation between surgical volume and mortality: an exploration of causal factors and alternative models.' *Medical Care,* 18(9): 940–959.

Marshall, M. Campbell, C. Hacker, J. Roland, M (2002). *Quality indicators for general practice.* London: Royal Society of Medicine.

McCain, C (2006). 'Using an FMEA in a service setting.' *Quality Progress,* 39(9): 24–29.

Miller, R West, C (2007). 'The value of electronic health records in community health centers: policy implications.' *Health Affairs,* 26(1): 206–214.

Morton, V. Torgerson, DJ (2005). 'Regression to the mean: treatment effect without the intervention.' *Journal of Evaluation in Clinical Practice*, 11(1): 59–65.

Persell, S. Wright, J. Thompson, J. Kmetik, K. Baker, D (2006). 'Assessing the validity of national quality measures for coronary artery disease using an electronic health record.' *Archives of Internal Medicine,* 166(20): 2272–2277.

Pogach, L. Engelgau, M. Aron, D (2007). 'Measuring progress toward achieving hemoglobin A1c goals in diabetes care: pass/fail or partial credit.' *Journal of the American Medical Association,* 297(5): 520–523.

Pringle, M (2002). 'Reflections on quality issues: quality becomes ever more complex.' *British Journal of General Practice,* 52(Suppl.): 47–48.

Reason, J (2000). 'Human error: models and management.' *British Medical Journal,* 320(7237): 768–770.

Rosen, AK. Reid, R. Broemeling, A-M. Rakovski, CC (2003). 'Applying a risk-adjustment framework to primary care: can we improve on existing measures?' *Annals of Family Medicine,* 1(1): 44–51.

Ross, L (1977). The intuitive psychologist and his shortcomings: distortions in the attribution process. In: Berkowitz, L (ed.). *Advances in experimental social psychology, Vol. 10.* New York: Academic Press.

Rouse, WB (2000). 'Managing complexity: disease control as a complex adaptive system.' *Information-Knowledge-Systems Management,* 2(2): 143–165.

Rouse, WB (2003). 'Engineering complex systems: implications for research in systems engineering. *IEEE Transactions on Systems, Man, and Cybernetics – Part C,* 33(2): 154–156.

Schneeweiss, S. Avorn, J (2005). 'A review of uses of health care utilization databases for epidemiologic research on therapeutics.' *Journal of Clinical Epidemiology,* 58(4): 323–337.

Seneta, E. Chen, JT. (2005). 'Simple stepwise tests of hypotheses and multiple comparisons.' *International Statistical Review*, 73(1): 21–34.

Simon, SR. McCarthy, ML. Kaushal, R. Jenter, CA. Volk, LA. Poon, EG. Yee, KC. Orav, EJ. Williams, DH. Bates, DW (2008). 'Electronic health records: which practices have them, and how are clinicians using them?' *Journal of Evaluation in Clinical Practice,* 14(1): 43–47.

Terris, DD. Litaker, DG (2008). 'Data quality bias: an under-recognized source of misclassification in pay for performance reporting?' *Quality Management in Health Care*, 17(1): 19–26.

Terris, DD. Litaker, DG. Koroukian, SM (2007). 'Health state information derived from secondary databases is affected by multiple sources of bias.' *Journal of Clinical Epidemiology,* 60(7): 734–741.

Tsai, AC. Vortruba, M. Bridges, JFP. Cebul, RD (2006). 'Overcoming bias in estimating the volume-outcome relationship.' *Health Services Research,* 41(1): 252–264.

Weiner, BJ. Shortell, SM. Alexander, J (1997). 'Promoting clinical involvement in hospital quality improvement efforts: the effects of top management, board, and physician leadership.' *Health Services Research,* 32(4): 491–510.

Zacharias, A. Schwann, TA. Riordan, CJ. Durham, SJ. Shah, A. Papadimos, TJ. Engoren, M. Habib, RH (2005). 'Is hospital procedure volume a reliable marker of quality for coronary artery bypass surgery? A comparison of risk and propensity-adjusted operative and midterm outcomes.' *Annals of Thoracic Surgery,* 79(6): 1961–1969.

Zyzanski, SJ. Flocke, SA. Dickinson, LM (2004). 'On the nature of analysis of clustered data.' *Annals of Family Medicine,* 2(3): 199–200.

3.4 *Using composite indicators to measure performance in health care*

MARIA GODDARD, ROWENA JACOBS

Introduction

Health-care performance is multi-dimensional and not easily captured by a single measure. Aspects of performance such as efficiency, quality, responsiveness, equity, outcomes and accessibility are all legitimate interests for the public and the policy-maker (Institute of Medicine 2001). It is not surprising therefore that there has been an explosion of interest in the generation, publication and interpretation of performance information in the health-care domain across the world, facilitated by the availability of information technology that allows for the capture of large amounts of complex data. This has occurred at all levels – whether individual practitioner, specific health services, health plans of provider organizations or entire health systems. However, the very abundance of such information can obscure users and policymakers' ability to make overall judgments about relative performance. Complex information presented over many dimensions may be difficult to comprehend and a lack of transparency presents opportunities for poor performance to go undetected. Users faced with multiple and disparate performance information will need to weigh the evidence and make trade-offs between different performance dimensions, thus increasing their processing burden. Some users may base decisions on a single performance dimension simply because it is the most clear. However, this will not necessarily be the most important.

In response to such issues, the use of summary or composite measures has become widespread in health and social policy arenas (Freudenberg 2003; Nardo et al. 2005). Such measures seek to combine disparate indicators of performance into a single score or index which can be used to compare (and sometimes rank) the relative performance of individuals, organizations or systems. This approach is not peculiar to health care; there are examples of the use of composite indicators

in many other sectors such as the environment, economy, technology, development, education and safety. It is also common practice to use composite measures to create league tables or rankings.

Composite indicators are in widespread use but their construction presents many methodological challenges. If not treated carefully and transparently these can leave them open to misinterpretation and potential manipulation. The accuracy, reliability and appropriateness of such indices need to be explored if major policy, financial and social decisions hinge on an organizations' performance as measured by composite indicators.

In this chapter we explore the advantages and disadvantages of constructing a composite indicator and describe the methodological choices made at each step in the construction. To illustrate these issues, we also describe some examples of current composite indicators in health care, highlighting good (and bad) practice in their development. We focus mainly on issues that are pertinent to the creation of composite measures rather than performance measurement in general, although of course there is much overlap.

Why use composite indicators to measure performance?

Composite indicators have a high profile in the media and play a potentially important role alongside the publication of individual performance indicators. However, they are not without drawbacks and any decision about the appropriateness of a composite measure will depend on a number of factors and the context in which they are to be used.

One of the main advantages of composite measures is that by focusing on a single measure they can give an overview of performance more readily than a plethora of diverse indicators. A single simple measure captures policy attention more easily and facilitates communication with the public about performance issues, thus enhancing public accountability. Composite measures also allow for the aggregation of a wide range of different types of performance data thereby ensuring that a rounded assessment of performance is presented rather than a focus on a single aspect. Comparison of single scores also means that it is easy to identify organizations that are performing poorly and should be priorities for improvement efforts.

On the other hand, composite indicators may lead to a number of dysfunctional consequences and there are several arguments against their use (Smith 2002). In particular, it is possible that a good composite score may mask serious shortcomings in some parts of a system. Transparency may be enhanced by summarizing performance but when performance is aggregated across a number of dimensions it may be difficult to determine the precise source of failings and therefore the remedial action required. In the health-care sector, data availability is often patchy across different domains and activities and therefore an indicator that is comprehensive in coverage is likely to rely on poor quality data along some dimensions. For example, outcome data are typically less readily available than process data and data on activity undertaken in the community are less accessible than those relating to secondary care. Conversely, unwanted behaviour can be induced by omitting measures for which data are unavailable as people focus only on what is measured.

The creation and publication of composite performance indicators can therefore generate both positive and negative outcomes, depending on the context in which they are used and the incentives they produce. The decision about whether composites are appropriate will always be a matter of judgment. However, where composites are used, the methodological choices made at each stage of construction will influence greatly their accuracy, reliability and appropriateness and have important implications for their impact. These include the choice of indicators; their transformation or standardization; the application of a system of weights; and the formation of the new composite. In the next section we provide some examples of the development and use of composite indicators in the health-care sector in order to illustrate issues arising from their construction and use.

Methodological issues and experience of using composite measures in health care

This section presents some of the methodological challenges that arise at each step of construction of a composite indicator. Where appropriate, these points are illustrated with discussions of composite measures of performance from health-care systems around the world.

Choosing units to assess and organizational objectives to encompass

These choices hinge on decisions about the boundaries of the units to be assessed and what aspects of performance these units will be held responsible for. They also depend on the target audience for the measures and the purpose of compiling the information. Measures of performance can be aggregated at a number of different levels – country, state, region, provider, health plan or physician. In addition, different elements of the health-care sector have overlapping boundaries – activities in one sector influence performance in another (e.g. primary care, secondary care, residential or long-term care and social services). Table 3.4.1 gives some examples of the coverage of composite indicator schemes.

Outside the health-care domain, many composite measures are reported at country level (e.g. environment, economic performance, quality of life). Within health, the WHO composite index of health system performance is probably the best known (WHO 2000). Despite much debate about the methodological detail, the publication of explicit rankings for 191 countries emphasized the potential power of using a single measure of performance to focus attention on important health-care issues. The Health Consumer Powerhouse has produced an annual health-care performance ranking for twenty-nine European countries (with recent addition of Canada) since 2005 (Health Consumer Powerhouse & Frontier Centre for Public Policy 2008).

The United States has produced composite measures of quality of care at state level for Medicare beneficiaries in fifty-two states, focusing on improvement as well as ratings. Jencks et al. (2000 & 2003) found that a state's average rank on the twenty-two indicators was highly stable over time with a correlation of 0.93 between the two periods. The better performing states appeared to be concentrated geographically in the northern and less populated regions (for both periods) but the geographical patterns of relative improvement by state were patchier.

Maclean's, a major mass-circulation magazine, publishes an annual health report that ranks Canadian regions according to their health-care performance. This is based on data published by the Canadian

Table 3.4.1 ***Examples of domains included in composite indicators***

Index	Organizations ranked	Domains
Commonwealth Fund National Scorecard*	States (United States)	Access Quality Potentially avoidable use of hospitals Costs of care Healthy lives
ECHCI	EU countries (+ Canada in 2007)	Patient rights/information Waiting times Outcomes Generosity Pharmaceutical coverage
Maclean's magazine	Regions (Canada)	Outcomes Resources Community health Elderly services Prenatal care Efficiencies
World Health Report	Countries (worldwide)	Health outcomes Inequality in health Fairness in financing Responsiveness Inequality in responsiveness
Healthcare Commission annual rating (2007 version)	Hospitals (England) + primary care trusts	Quality of services Use of resources
Healthcare Commission star ratings (prior to 2005)	Hospitals (England)	Key target areas (e.g. waiting times, finance) Clinical focus Staff focus Patient focus

*Gives disaggregated results rather than a composite indicator but produces overall rankings

Institute for Health Information in a series of annual reports and a series of health indicators for the sixty-three largest regions, covering 90% of the population (Canadian Institute for Health Information 2001, 2001a & 2007). In the 2001 report, the composite performance scores ranged from 89.5 in North/West Vancouver, British Columbia to 73.4 in North Bay/Huntsville, Ontario.

Composite measures are created most commonly at provider level, usually a hospital. This focus is understandable because it is easier to see a direct line of accountability between the performance of that organization and the hospital management than (say) from the state, region or country downwards. The United States produces vast amounts of performance information; composite measures of performance have been constructed for hospitals and nursing homes for some time. For example, HealthGrades gives detailed performance information for consumers, providers and health plans (http://www.healthgrades.com). This organization gathers together a wide variety of information (e.g. Medicare inpatient data; range of specialized information provided by states) to provide detailed profile information on hospitals; star ratings (from one to five) for ten clinical areas; and (based on these individual star ratings) an overall ranking of the top fifty best hospitals. America's Best Hospitals guide (www.rti.org/page.cfm?objectid=EDFAA2A9-4725-488E-83AE91A9442C9727) has operated for over fifteen years and is reported widely in the American press. This provider-level system ranks hospitals in sixteen specialties and by their overall performance. Hospitals that score at or near the top for a minimum of six specialties are classified as super elite.

In England, hospital trusts have been the focus of composite ratings for some time – the star ratings. A composite index score for each NHS organization places them in one of four categories: from three stars (highest levels of performance) to zero stars (poorest levels of performance). At the outset in 2001 only acute trusts were included (Department of Health 2001); specialist trusts, ambulance trusts and indicative ratings for mental health trusts were added later (Department of Health 2002). By 2003, all NHS providers were covered, including local purchasers of health care (primary care trusts). Further indicators have been published every year since but the nature of the performance assessment has altered over time and now there is less emphasis on summary measures (Healthcare Commission 2004, 2005 & 2007).

There are also composite measures for specialties such as paediatrics, cardiac surgery, long-term care and chronic conditions. At physician level, many different incentive schemes are based on linking income with performance but not all use a single composite score to measure performance. In New York, a demonstration project linked physician payment to performance on a composite compiled from process and outcome data for diabetes care (Beaulieu & Horrigan 2005).

As illustrated above, much of the measurement activity at national level has taken place in the acute hospital setting; even the star ratings for English primary care trusts were dominated by health-care activity in the secondary sector. There have been examples of composite indicators at primary-care level e.g. the Summary Quality Index (SQUID) in England (Nietert et al. 2007). These may be useful locally but tend not to have a national profile.

Choosing the indicators

This is probably one of the most important steps. Careful judgment is required as effort will be focused on the included indicators, potentially at the expense of achievement on those excluded.

Data availability

In practice, many composites are often opportunistic and incomplete (measuring aspects of performance captured in existing data) or are based on highly questionable sources of data. Either weakness can seriously damage the credibility of the composite (Smith 2002). The choice of indicators is most often constrained by data availability and thus may give an unbalanced picture of health services. The excluded indicators may be equally (or more) important but simply more difficult to measure.

The higher the level at which composites are created and the broader their scope the greater the issues of data availability and lack of comparability. The WHO composite index of health system performance was produced for 191 countries and sought to be comprehensive in coverage. It measured five domains: (i) overall health outcomes; (ii) inequality in health; (iii) fairness of financing; (iv) overall health system responsiveness; and (v) inequality in health system responsiveness. Much of the debate about the index has focused on appropriateness of the measures used to capture these domains and the source and

robustness of the data (e.g. Almeida et al. 2001; Appleby & Street 2001; Navarro 2002; Nord 2002; Smith 2002; Williams 2001).

The Euro-Canada Health Consumer Index (ECHCI) aims to cover issues of relevance to the consumer and therefore focuses on five areas: (i) patient rights/information; (ii) waiting times; (iii) outcomes; (iv) generosity (activity rates); and (v) pharmaceuticals (e.g. access to new drugs, subsidies). A total of twenty-seven indicators were included in their most recent index but it was noted that the original, larger set had been pared down due to lack of data (Health Consumer Powerhouse & Frontier Centre for Public Policy 2008). It is clear that there will be a trade-off between an ambitious aim of deriving a composite measure, capturing complex and comprehensive health performance dimensions for a wide range of countries, and the practical issues of gathering good data on such dimensions.

The availability of data explains partly why most performance measures focus on hospital rather than community services. However, even within a sector there are many choices about the areas to be covered. For example, there has been criticism of the Canadian ratings of regions for excluding psychiatric care and the English star ratings for relying on process measures and focusing solely on indicators for which there are national targets. Also, many systems rely on indicators in only a few key disease areas. For example, the American state-level indicators for Medicare beneficiaries (Jencks et al. 2000 & 2003) cover six clinical areas: (i) acute myocardial infarction (six indicators); (ii) heart failure (two); (iii) stroke (three); (iv) pneumonia (seven); (v) breast cancer (one); and (vi) diabetes (three). The choice of indicators tends to over-represent inpatient and preventive services and under-represent ambulatory care and interventional procedures. However, an explicit rationale informed the selection of clinical areas according to the following criteria:

- disease is a major source of morbidity or mortality
- certain processes of care are known to improve outcomes
- measurement of these processes is feasible
- offers substantial scope for improvement in performance
- managerial intervention can potentially improve performance.

Lack of agreement about the data definitions and lack of consistency in interpreting the measures can also lead to partial representation of performance. The Canadian regional-level composites have been criti-

cized on this basis but it has been noted that the number of indicators has expanded over time and new data have been incorporated as they become available (e.g. stroke survival in the latest round). In addition, this has prompted the quest for improvements in data quality. For example, only a handful of regions were able to provide waiting time information because of variations in definitions and collection methods. This will be addressed in future. The Canadian Institute for Health Information (http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=home_e) notes that no comparable data were available for the public and providers five years ago so the rankings represent significant progress, despite the gaps in coverage. Improvements in data quality and availability may be a positive side-effect of attempts to create such indicators.

Data availability aside, the choice of indicators may reflect political priorities for performance. For example, the early stages of the English star ratings were dominated by waiting times and financial issues. Other indicators were included but given less weight in the final performance rating.

Type of indicators

There has been much debate about the pros and cons of different types of performance indicators in health care, particularly process and outcome measures (see Chapter 5.5*)*. A focus on outcomes directs attention towards the patient (rather than the services provided by the organization). However, there can seldom be any confidence that outcome measures such as current health status are indicators of current health system performance. For example, it is clearly impractical to wait for some health outcomes (that may take years to emerge) before making a judgment on performance. Furthermore, the collection of outcome data may impose high costs on the health system. Finally, there are issues around attribution and the extent to which health status can be attributed solely to the health-care system *(*see Chapter 3.3*)*. In such circumstances, it becomes necessary to rely on measures of health system process rather than health status outcome.

Process measures can be more meaningful for some users of performance ratings. For example, the SQUID composite measure of quality of care in primary care in England was created by combining thirty-six process and outcome measures (Nietert et al. 2007). More than one hundred ambulatory-care practices receive quarterly data on the

patient level (proportion of recommended care received) and practice level SQUIDs (average proportion of recommended care received by the practice's patients). Measures of recommended care relate to indicators such as the proportion of the target population receiving specific interventions or tests (e.g. beta blockers, screening tests, counselling). The authors note that, unlike many composite measures, their SQUID score has a meaningful clinical interpretation which probably accounts for its acceptability to doctors.

Patients are becoming increasingly vocal in demanding that health care should be responsive to concerns over and above the health outcomes that result from treatments. This concern with the patient experience covers issues as diverse as promptness, autonomy, empowerment, privacy and choice *(*see Chapter 2.5*)*. Such performance measures may be particularly appropriate when there are large variations in the responsiveness of organizations, as indicated by hospital waiting times in many publicly funded health systems. The WHO ratings of health-care systems included a measure of responsiveness to citizens as this was thought to be an important element in the health-care experience and one which might vary considerably between systems. The ECHCI is aimed at consumers and therefore many of the indicators relate to process issues of relevance to their audience, such as waiting times and the availability of a wide range of information via different media. The English performance ratings now include measures of patient satisfaction taken from annual surveys. In some circumstances (e.g. management of chronic diseases) process measures will be far more relevant to patients than outcome measures (Crombie & Davies 1998).

Collinearity between indicators

The final issue relating to the choice of indicator concerns the potential for performance indicators that measure similar aspects of performance to be highly correlated with each other. The concern is that the inclusion of variables which are highly collinear will effectively introduce some sort of double counting. It has therefore been argued that a chosen set of indicators should be reduced by selecting between indicators with high correlations. This may be desirable for reasons such as parsimony and transparency.

Multivariate statistical methods are available to investigate relationships between the indicators within a composite. Principal components analysis (PCA) and factor analysis (FA) may be used to extract statistical correlations between indicators to enable identification of a core group of indicators that statistically best represent the remaining excluded indicators (Joint Research Centre 2002). Factor analysis of individual measures used in two major performance schemes in the USA (HEDIS, CAHPS®) have frequently illustrated that it is feasible to achieve parsimony by aggregating indicators into one or a small number of composites. For example, for CAHPS, six out of thirty-three factors provided the best description of variation at patient level and three out of thirty-three explained much of the variation at hospital level (O'Malley et al. 2005). For HEDIS, a single composite explained 38% of the variation at hospital level and the use of three composites improved this to 60% (Lied et al. 2002). Similar analysis using a combination of all indicators in HEDIS and CAHPS illustrated that they could be separated into a four-factor solution that explained 64% of the variation in the measures (Zaslavsky et al. 2002). Other composite measures have been created from variables found to be generally uncorrelated with each other e.g. quality in cardiac surgery (O'Brien et al. 2007).

If statistical techniques are used to choose the variables for inclusion, it is likely that highly collinear variables will be excluded through model specification tests for multicollinearity. The choice of one variable over an alternative highly collinear variable may not alter rankings greatly but may affect the judgments on a small number of units, with extraordinary performance in either of those dimensions. It may therefore be subject to dispute and challenge.

Combining indicators to create a composite

The next stage is to aggregate the chosen indicators that are likely to be measured in different units and on different scales. Aggregation needs to be undertaken in a consistent manner in order to ensure that the composite measure produced is easily understood and has the intended incentive effects. The combination of the measurement scale used for individual indicators, and the weights applied to add them

together, can affect the interpretation of changes in the composite indicator. The aim is to be transparent about how much improvement is required in one constituent indicator to compensate for deterioration in another.

Three key steps in aggregation are described below: (i) transformation of individual indicators; (ii) weighting; and (iii) application of decision rules.

Transformation of individual indicators

Transformation is less important if it is possible to specify a weight that indicates the relative value to the composite of an extra unit of attainment in that dimension at *all* levels of attainment. However, most indicators that make up a composite will be non-linear – an x-point change of the variable on one part of the scale will have a completely different effect on assessed performance than an x-point change on another. This requires them to be transformed in some way to enable aggregation into a composite. Other reasons for transformation include the need to allow for extreme values (outliers) which may otherwise skew the composite and the desire to add together indicators measured in different units.

A number of methods are available for transforming the underlying indicators including ranking, normalizing, re-scaling, generating various types of ratio variables, logarithmic transformation or transforming variables to a categorical scale. All of these can impact on the final outcome of the composite indicator. Table 3.4.2 shows some examples of the impact of choice of transformation method using hypothetical data for ten organizations. The methods have been surveyed elsewhere (Nardo et al. 2005) but not one model fits every set of circumstances – each is associated with pros and cons.

It is useful to explore how alternative measures for standardization impact on final performance rankings. For example, Lun et al. (2006) show that the use of Z scores (use the mean and standard deviation to adjust raw scores) rather than raw scores, dramatically changes the ranking of quality of life for 103 Italian provinces, with some moving 88 places in the ranking. This method gives greater weight to variables with extreme outliers. The use of Min-Max methods (use the differences between minimum and maximum scores) gives less weight to outliers but also changes rankings substantially. Similarly, Cherchye

Table 3.4.2 *Examples of impact of different transformation methods*

Unit	Raw data	Ranking	Standard-izing	Rescaling (best = 100, worst = 0)	Difference from mean (mean = 100)	Difference from leader (best = 100)	Threshold above and below mean (threshold = 20%)	Logarithmic	Categorical scale (percentiles = 0.75, 0.5 and 0.25)
Unit 1	2.85	1	2.01	100.00	174	100.00	1.37	0.45	3
Unit 2	2.08	2	1.09	71.01	94	71.93	0.65	0.31	3
Unit 3	1.58	3	0.54	53.99	47	55.44	0.22	0.20	3
Unit 4	1.35	4	0.28	45.65	24	47.37	0.02	0.13	2
Unit 5	1.03	5	-0.09	34.06	-8	36.14	-0.27	0.01	2
Unit 6	0.86	6	-0.29	27.90	-25	30.18	-0.43	-0.07	1
Unit 7	0.59	7	-0.60	18.12	-52	20.70	-0.67	-0.23	1
Unit 8	0.43	8	-0.79	12.32	-68	15.09	-0.81	-0.37	0
Unit 9	0.28	9	-0.96	6.88	-83	9.82	-0.95	-0.55	0
Unit 10	0.09	10	-1.18	0.00	-102	3.16	-1.12	-1.05	0

Unweighted average = 1.11
Standard deviation = 0.86

et al. (2007) illustrate the hypothetical impact of varying methods of normalization for country rankings and also question the wisdom of making statements about the resulting normalized scores e.g. that the global performance of organization/country X is 5% better than that of organization/country Y.

The choice of an appropriate method of transformation is therefore dependent on both the nature of the indicators and the composite's desired incentive effects on performance. For instance, it may be appropriate to allow extreme values on some indicators to influence overall performance on the composite when the intention is to reward exceptional behaviour on a few indicators, rather than average performance on all.

Weighting

In order to achieve a specific final score on the composite measure, the efforts required to improve performance on a sub-indicator will depend on the weight applied to it. The incentive effects of weighting are therefore potentially very powerful – the ranking of a particular organization can change dramatically if an indicator on which the organization excels or fails is given more weight. A weight indicates the relative opportunity cost of achieving each of the underlying indicators; it can be designed to equalize this across all indicators or to put more emphasis on some at the expense of others. This represents a trade-off in the efforts to achieve good performance on each indicator.

Differential weights are chosen for a variety of reasons although the usual interpretation is to reflect the importance of the underlying indicators (Cherchye et al. 2007). However, there should be consideration of the interaction between the way in which the indicators have been transformed (see above) and aggregated and the weights subsequently applied. In particular, in most methods of aggregation weights represent the trade-off between indicators. This suggests that it is acceptable for good performance in one domain to be offset by poor performance in another. However, if weights are meant to reflect relative importance then alternative methods of aggregation that do not allow for such compensatory behaviour must be used.

The impact of choices has been illustrated using health performance data from England where varying weights have been shown to have a

Table 3.4.3 ***Examples of methods to determine weights***

Statistical approaches	Factor analysis
	Principal components analysis
	Data envelopment analysis
	Benefit of the doubt
Participatory approaches	Budget allocation
	Analytic hierarchy process
	Conjoint analysis
	Opinion polls and surveys

major effect on rankings (Jacobs et al. 2005). Also, it is observed that a region such as Edmonton can rate near the bottom of the rankings for low birth weight infants but still emerge at the top of the overall ranking within their group due to the combined impact of the complex set of weights used in the Canadian system (Page & Cramer 2001). Weights may also be chosen to reflect other characteristics of the indicators – for instance, those which have more reliable underlying data may be given greater weight in the final indicator. However, this may reinforce the dependence on easily measured and available data within performance results (Freudenberg 2003; Nardo et al. 2005).

Having decided on the purpose of the weighting system, the weights have to be derived. This can be achieved by using either a range of statistical techniques or participatory techniques that generally employ the judgment of individuals. Some of the relevant techniques for determining weights are listed in Table 3.4.3. The use of participatory methods involves fundamental consideration of the preferences used in the elicitation of those weights – whether those of policy-makers, providers, purchasers, patients or the public. The weights used will usually reflect a single set of preferences but the preferences of policy-makers, individual providers and the broader public are likely to vary.

Participatory techniques include direct interviews, surveys and public opinion polls. More advanced techniques enable the analyst to elicit trade-offs between several attributes or performance dimensions. These include the analytic hierarchy process (AHP) in which opinions are systematically extracted by a pair-wise comparison between different dimensions or attributes of performance (Saaty 1987). Conjoint analysis also has been used widely in the health-care context (Ryan &

Farrar 2000). This attempts to elicit values and trade-offs between the various attributes of a good service or, in this context, aspects of performance. Both approaches are able to deal with multiple attributes, particularly helpful in the context of health where there is likely to be interest in a wide range of dimensions of performance.

Three different approaches to eliciting preferences are illustrated by a British experiment organized by a television company; the WHO country performance rankings; and America's Best Hospitals in the United States. In 2000, a polling organization surveyed 2000 people across England, Scotland and Wales to obtain their preferences for selected aspects of health authority performance (Appleby & Mulligan 2000). Three methods were used to elicit preferences: (i) ranking from most to least desired indicator; (ii) budget-pie –respondents were asked to allocate a 'budget' of sixty chips between six performance indicators; and (iii) conjoint analysis. This offered the advantage of multi-attribute approaches as well as considering simpler trade-off methods. The authors spent considerable efforts to ensure that their weighting system reflected variations in views obtained from the different methods.

In contrast, the weighting system underlying much of the WHO rankings depended upon expert opinion. Dimensions of responsiveness were scored by around 2000 key informants from 35 countries who answered questions about their own countries and were then asked to score responsiveness as a whole. Another group of 1000 people ranked the 7 aspects of responsiveness in order of importance in a web-based exercise; weights were assigned based on the rankings. Mean scores on each aspect were multiplied by weights and summed to give an overall responsiveness score. The final dimension (equity in responsiveness) was calculated by asking informants to make judgments about the subgroups that they thought were treated with less responsiveness. Scores were assigned to subgroups based on the number of times that they were mentioned by country informants, multiplied by that group's share in the population. The products were summed and transformed to give an overall score. Finally, the individual scores on five dimensions of performance (including the responsiveness measure discussed above) were aggregated to create an overall attainment score. Individual measures were transformed to a 0–100 scale and summed using weights of either 0.25 or 0.125, based on the views of about 1000 people from 123 countries, half of whom

were WHO staff. There has been widespread debate about the pros and cons of the approaches used (e.g. Almeida et al. 2001; Williams 2001).

America's Best Hospitals in the United States is another example of the use of expert opinion. This is based on survey responses and uses reputation as one of three dimensions of the composite indicator. A random sample of specialists (in each specialty) is asked to list the five best hospitals for 'difficult' cases in their specialty. This is undertaken without reference to geography or costs.

The use of statistical or empirical methods (rather than preferences) to create weights might be expected to raise fewer issues but the methodological challenges are still substantial (e.g. Lun et al. 2006). If it is possible to demonstrate that alternative approaches have little impact then this will help to build confidence in the results. For example, Zaslavasky et al. (2002) used three alternative statistical approaches to create weights for health performance and found similar final results (Mullen & Spurgeon 2000).

An entirely different approach uses data envelopment analysis (DEA) to create performance ratings without the need to incorporate fixed weighting systems. This is sometimes called the benefit of the doubt approach in the context of performance ratings. Essentially, this allows the use of flexible weights that vary across domains and between the organizations being assessed (Cherchye et al. 2007). For example, the weights assigned to different dimensions of performance for a country are derived from the country data. The core idea is that the country's good relative performance on a particular sub-indicator signals that the indicator has policy importance in that country and hence should be assigned a higher weight than in another country where relative performance on that dimension is weak. It is not possible to document all the pros and cons of such an approach (see Cherchye et al. for details) but one main drawback from a policy perspective is that the results may be difficult to reconcile with general views on the relative importance of different aspects of performance. For example, an organization may be excellent at a dimension of performance that is considered rather marginal in the overall healthcare context and it may seem inappropriate if their final composite score is influenced heavily by their performance along that dimension. This can be addressed to some extent through the use of restrictions – limiting the share of the total composite result that can be gained from

specific sub-indicators. This can be achieved in several ways (depending on the strength of consensus about the importance of different indicators) and allows a great deal of flexibility in assigning weights using the revealed performance of organizations. This approach is probably of most value where the aim is to combine very disparate indicators at a high (e.g. country) level, where relative performance will be affected heavily by a wide range of factors.

In conclusion, there appears to be little consensus about the preferred technique for participatory methods (Dolan et al. 1996) and it is likely that the different methods will lead to the emergence of different preference sets. These examples illustrate the difficulties in eliciting preferences and devising weights and serve as reminders that a composite cannot be presented as 'objective' (Smith 2002). The choices about who and how to ask depend in part on the nature of the performance domains to be captured. Where responses require a great deal of technical or background knowledge it is legitimate to target experts, although the definition of expert may be controversial. For example, it could be argued that WHO staff may not necessarily have more knowledge than ordinary members of the public in some areas of questioning. In a complex area such as health care, multi-attribute approaches may be preferable to more simplistic methods. The former are more expensive to organize and are feasible only where a fairly limited set of domains is considered, otherwise the comparisons become too unwieldy. In all cases, comparisons between countries present particular challenges for ensuring consistency in elicitation methods. Statistical methods offer an alternative and may be especially valuable where high-level performance across countries is being considered. However, these can be difficult to explain and are less intuitive for the public and policy-makers than participatory approaches.

Application of decision rules

Rather than attaching explicit weights to transformed indicators, a set of decision rules can be applied to produce a composite indicator. Such rules reflect views on the importance of achieving certain standards. They set the boundaries within which performance scores will be allocated (e.g. defining what constitutes a good or poor score on an indicator); or they may disallow a good performance score if an organization fails to meet a particular target on a single indicator.

The rules are often applied sequentially and implicitly introduce a set of weights.

One example of this was the construction of the scorecard for acute hospitals in the star ratings system in England. This applied a complicated algorithm with a set of sequential decision rules to determine the ultimate star rating (composite indicator). The star ratings for trusts comprised four areas: (i) key government targets; (ii) clinical focus; (iii) patient focus; and (iv) capacity and capability. The key government targets were the most significant factors in determining overall performance ratings. Performance was assessed in terms of whether the target had been achieved; whether there was some degree of underachievement; or whether the target was significantly underachieved (threshold type variables). The methodology broadly entailed transforming the underlying key targets and performance indicators into categorical variables of either three or five categories. The performance indicators in the patient, clinical, and capacity and capability focus areas were categorized into one of five performance bands (from five points for the best performance to one for the worst). The thresholds for deciding the cut-offs were not necessarily the same for each variable and individual band scores were combined to produce an overall score per area. All indicators were weighted equally within their scorecard area to ensure that each scorecard area carried the same weight, despite differing numbers of indicators. A complex six-step process imposed a sequential set of decisions on achievement on the various key variables to determine the final star rating. Evidence suggests that the application of such rules and subtle changes to their application can be hugely influential in the final outcome of the composite measure – small changes in decision rules can move hospitals from one end of the performance league table to the other (Jacobs et al. 2005). Reeves et al's (2007) comparison of five different methods of combining clinical quality indicators at primary-care provider level shows that the rules applied to the scoring of sub-indicators can change rankings dramatically. The pros and cons of using rules that set thresholds rather than dichotomous measures have also been analysed (Aron et al. 2007).

O'Brien et al. (2007) illustrate the impact of different approaches to aggregation on a composite score of provider ratings for quality in cardiac surgery. Their analysis investigated a wide set of options for combining eleven indicators of quality within and across four

domains of care. They used data from over 133 000 procedures to test out methods such as scoring, scaling, opportunity-based approaches, latent variable models and all-or-nothing rules. In contrast to the analysis of the English data reported above, they concluded that 'inferences about a provider's quality were robust and largely insensitive to choice of methodology' (O'Brien et al. 2007, p. S21). However, they rejected some approaches (e.g. use of literature or expert views to assign importance weights to measures) and their range of measures was probably less diverse. O'Brien et al. focused on narrow definitions of quality for one specific type of care while the English system covered financial, clinical quality, staffing and other dimensions at the whole hospital level.

Sometimes the application of rules can produce a lack of transparency but there are often good reasons for such an approach. In particular, they can ensure that certain minimum requirements are met. For instance, O'Brien et al's (2007) analysis uses an all-or-nothing rule for some dimensions of quality – hospitals that do not *fully* meet the stated standard receive a zero score on that dimension with no credit for partial compliance (e.g. 100% of patients must receive the stated quality of care; 100% of patients must avoid complications). These approaches are common when it is felt appropriate to set a high benchmark on a particular domain of performance. Decision rules to attain minimum standards may be particularly pertinent for a hospital accreditation process. They are also useful stepping stones in performance reward systems where a baseline level of reward is contingent on attaining minimum standards in key areas and less stringent requirements are placed on other dimensions.

Interpretation and use of composite indicators

A composite indicator derived from a number of sub-indicators has the potential drawback that the indicators themselves will be subject to some degree of uncertainty. If they are combined into one composite without due regard for the underlying distribution of the variables their results may lack robustness. There are various methods for investigating the nature of the sub-indicators. Much research has been undertaken to look at the features of available sub-indicator data in terms of their appropriateness for incorporation into a composite performance measure – for example, looking at the extent of miss-

ing data, variability in performance, coverage of the relevant patient population, predictive properties of a process indicator etc. This has been undertaken in many different contexts e.g. paediatrics (Bethell et al. 2004) and nursing-home care (Berg et al. 2002). A more detailed approach attempts to separate out random fluctuations in the underlying variables from those attributable to actual differences in performance and to create confidence intervals around the resulting scores. Jacobs et al. (2005) explored this using English data and employing Monte Carlo simulation methods in order to demonstrate that there was a small group of providers who could – with confidence – be said to be performing better or worse than others but that such statements were less feasible for many in the middle ranks. Similarly, the authors of an analysis of Italian quality of life data were able to demonstrate some coarse groups of differentially performing provinces (Lun et al. 2006).

Another problem arises in interpretation – composite scores often feed into performance rankings and will produce conflicting results if slightly different composites are used. As illustrated earlier, small changes in methods can affect the resulting composites, even if similar data are used. Different data sources can cause even more confusion. This may be similar to the conflicting results that arise on individual indicators over a range of performance measures (when there are large variations in organizational rankings) but the conflicts are far more visible and stark and more likely to capture public interest. For example, several schemes in the United States receive a great deal of consumer attention but are constructed in slightly different ways. HealthGrade's composite scores for clinical areas are used to produce an overall ranking of the top fifty best hospitals. America's Best Hospitals ranks hospitals in sixteen specialties and by overall performance (US News & World Report 2007). Ratings are based on three areas: (i) reputation; (ii) mortality; and (iii) range of factors such as accreditation scores, inputs, availability of technology. The three elements are combined with equal weights and hospitals are ranked within each specialty. Hospitals that score at or near the top of the rank for a minimum of six specialties are classified as super elite. The Centers for Medicare and Medicaid Services (CMS) launched Hospital Compare in 2005 in order to provide patients with information on hospital quality, rather than targeting providers or regulators (www.cahps.ahrq.gov/content/products/HOSP/PROD_HOSP_Intro.asp). Data from 4000 hospitals are used to compile quality indicators.

Results from America's Best Hospitals and Hospital Compare have been compared in order to explore the consistency between rankings (Halasyamani & Davis 2007). Hospital Compare does not produce rankings using composite scores but its core performance measures were used to examine quality in three areas: acute myocardial infarction, congestive heart failure and community-acquired pneumonia. The scores were combined with equal weights to produce rankings of the hospitals. The properties of the indicators within each group were examined for statistical robustness and Hospital Compare scores were calculated for the hospitals included in America's Best Hospitals' rankings – for heart and heart surgery; respiratory disorders; and overall quality (roll of honour hospitals). The authors found that the separate measures for the three clinical areas had good internal consistency but there was little agreement between the Hospital Compare scores and America's Best Hospitals' ranks. Indeed, several of the 'best' hospitals scored below the national median in the disease area scores. There are reasonable explanations for some of the disparities – for instance, America's Best Hospitals relies heavily on mortality rates and on physicians' perceptions of reputation; Hospital Compare looks at delivery of disease-specific evidence-based practices. However, the analysis illustrates the difficulties of relying on a composite measure and ranking without adequate reflection on the nature of the underlying indicators.

Similarly, analysis of the HealthGrades rankings of hospitals has shown that these produce groups of hospitals that differ in the quality of care but do not differentiate well between any two hospitals' individual mortality rates. The authors claim that hospital performance is thus seriously misrepresented to the public (Krumholz et al. 2002). Similar results have been found by others (Werner & Bradlow 2006). Analysis of the rankings of cardiac hospitals produced by a national newspaper in the United States concluded that many of the newspaper's top-fifty hospitals were indeed performing significantly better than their peers but some were failing to provide evidence-based best practice. Also, some lesser-rated hospitals were in fact routinely providing cardiac care that accorded with national guidelines (Williams et al. 2006). It is debatable whether the public can be expected to appreciate the differences in scope and methodology and draw appropriate conclusions.

When incentives are attached to performance results, their accurate interpretation and robustness becomes even more vital. In the early days of the English star ratings much discontent was voiced at their use as a means of rewarding and penalizing managers – hospitals that obtained a three star rating for a consecutive number of years could apply for foundation status which confers significant financial and managerial decision-making freedoms and autonomy from central involvement (Cutler 2002; Kmietowicz 2003; Miller 2002; Snelling 2003). However, star ratings varied from year to year; in some extreme cases hospitals fell from three stars to zero stars within one year. These shifts seldom reflected dramatic changes in overall performance and usually were due to the application of varying decision rules that blocked a high overall score if hospitals fell below a minimum standard in one indicator. The Healthcare Commission subsequently broadened performance assessment in order to focus less on a composite score and more on a whole range of performance indicators (Healthcare Commission 2007).

The United States has been at the forefront of attaching financial incentives to performance ratings in health care. In July 2003, Premier (a nationwide organization of not-for-profit hospitals) and the CMS launched the Hospital Quality Incentive Demonstration Project (HQID) (Premier 2005; Centers for Medicare and Medicaid Services 2005) – the pay-for-performance scheme. CMS rewards participating hospitals that achieve superior performance by increasing their payment for Medicare patients. The project covers five clinical areas and hospital performance for each is aggregated into a composite score to establish baseline performance. Each composite consists of a process score (twenty-seven indicators) and outcome score component (seven indicators) weighted proportionally to the number of each type of indicator in the category. The composite *process* score in each category is created by summing the numerator and denominator values for each indicator and then dividing the totals. The composite *outcome* score in each category is created by generating a survival index of actual divided by expected survival rate. Each is then multiplied by the component weighting factor. The composite score is used to identify the hospitals eligible for incentive payments. Those in the top decile of quality for a given clinical area receive a 2% bonus of their Medicare payments for the given condition; hospitals in the second decile receive

a 1% bonus. Composite quality scores are calculated annually. In year three, payments are adjusted for those hospitals that do not achieve performance improvements above baseline.

There has been much discussion about the impact of pay for performance. This is difficult to evaluate given the plethora of published quality ratings which may go some way towards encouraging performance improvement, even in the absence of financial incentives. A recent evaluation of composite measures compared public reporting of performance alone (through Hospital Compare ratings) and the pay-for-performance scheme and was able to make more relevant comparisons by careful matching of participating and excluded hospitals. This indicated that the incremental effect of financial incentives attached to the composite measures was between 2.6% and 4.1% (Lindenauer et al. 2007).

Conclusions

The use of composite measures of performance is common in many countries and sectors. Many of the technical and methodological issues associated with the construction of composites are similar to those faced in the general field of performance measurement and are not unique to the context of composite measures. However, in this chapter we have focused on some of the key issues that are particularly pertinent when attempting to combine indicators – mainly issues related to the choice of sub-indicators; the nature of their transformation; weighting schemes and decision rules; and the interpretation and use of composite scores and rankings. We have demonstrated that choices are made at each stage of their construction, often based on practical considerations such as data availability. These may appear largely technical or of minor significance but in fact can have a fundamental impact on the final performance results. This may call into question the utility of composite scores but it is hoped that the publication of composite measures can also lead to greater attention to issues of data quality and comparability and a search for a more satisfactory methodology.

Some recent moves have aimed to reduce reliance on composites alone. For example, in England the Healthcare Commission incorporated the overall ratings for providers (now designated as 'excellent', 'poor' etc) into a broader assessment process which contains

a plethora of information (Healthcare Commission 2007). Dr Foster Intelligence (an independent organization set up to publish performance data) recently decided not to publish best-hospital rankings but to present a limited number of league tables based on single measures and selective reporting of other dimensions of performance (Dr Foster Intelligence 2007). In the United States, the Commonwealth Fund National Scorecard ranks states' overall performance across five dimensions but this is published alongside the detailed results and rankings disaggregated for all thirty-two indicators (Commonwealth Fund 2007).

An array of performance data can offer some advantages but we argue that composite scores play an important role in helping to focus attention on key aspects of performance in a way that the public can understand easily. They are therefore an important means of promoting accountability and providing the public with useful information about physicians, provider organizations and their overall health-care systems. Composite scores allow the best performers to be recognized easily and indicate those that need to improve. They can offer some flexibility at a local level if there is scope for managers to improve in their own priority performance domains and to make efforts where they will secure the most overall gain in performance.

Our main recommendation for policy-makers is to make methodological decisions explicit and at each stage to undertake detailed exploration of the nature of the underlying indicators and the final scores' sensitivity to the decisions to be made. Misleading results may result from underestimating the impact of what appear to be just technical decisions. The conceptual limits of composite indicators should be borne in mind and published with explanations of the choice of indicators, the transformation method and the weighting structure. Consideration should also be given to demonstrating the confidence intervals surrounding composite scores although it is a challenge to do this in a user-friendly way. Publication of the disaggregated data that underpin the composite or publication of additional supplementary data alongside the composite results may be a useful compromise as long as this does not obscure entirely the purpose of providing a concise summary of performance. Explanations of the limits of the composite may help interpretation and transparency by clarifying what policy objectives are being maximized. Composite measures are amenable to being linked with incentive mechanisms for good performance but

powerful financial and other incentives should not be used unless there is confidence in the way in which the composites have been derived.

The creation of league tables and rankings is often one of the main purposes behind the construction of composite indicators as they facilitate easy comparisons. Such tables enjoy a high profile in the popular press and make very attractive headlines, especially when targeting the 'worst' performers. There is a danger that health-care organizations can be damaged by premature or inaccurate publication of such information without adequate accompanying health warnings. However, as long as there is open discussion of the processes by which they are derived and some careful interpretation then publication in this format may be an important first step in revealing important performance variations which might otherwise go undetected, unreported and unaddressed.

References

Almeida, C. Braveman, P. Gold, MR. Szwarcwald, CL. Ribeiro, JM. Miglionico, A. Millar, JS. Porto, S. Costa, NR. Rubio, VO. Segall, M. Starfield, B. Travessos, C. Uga, A. Valente, J. Viacava, F (2001). 'Methodological concerns and recommendations on policy consequences of *The world health report 2000*.' *Lancet,* 357(9269): 1692–1697.

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

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

Aron, D. Rajan, M. Pogach, L (2007). 'Summary measures of quality of diabetes care: comparison of continuous weighted performance measurement and dichotomous thresholds.' *International Journal for Quality in Health Care,* 19(1): 29–36.

Beaulieu, N. Horrigan, D (2005). 'Putting smart money to work for quality improvement.' *Health Services Research*, 40(5 Pt. 1): 1318–1334.

Berg, K. Mor, V. Morris, J. Murphy, K. Moore, T. Harris, Y (2002). 'Identification and evaluation of existing nursing home quality indicators.' *Health Care Financing Review,* 23(4): 19–36.

Bethell, C. Peck Reuland, C. Halfon, N. Edward, L (2004). 'Measuring the quality of preventive and developmental services for young children.' *Paediatrics,* 113(Suppl. 6): 1973–1983.

Centers for Medicare and Medicaid Services (2005). *Medicare 'pay for performance (P4P)' initiatives.* Baltimore, MD: Centres for Medicare and Medicaid Services (http://www.cms.hhs.gov/apps/media/press/release.asp?counter=1343).

Cherchye, L. Moesen, W. Rogge, N. Van Puyenbroeck, T (2007). 'An introduction to 'benefit of the doubt' composite indicators.' *Social Indicators Research*, 82(1): 111–145.

CIHI (2001). *Health care in Canada 2001: a second annual report*. Ottawa: Canadian Institute for Health Information.

CIHI (2001a). *Health indicators 2001*. Ottawa: Canadian Institute for Health Information.

CIHI (2007). *Health indicators 2007*. Ottawa: Canadian Institute for Health Information.

Commonwealth Fund Commission on a High Performance Health System (2007). *Aiming higher: results from a state scorecard on health system performance*. New York: The Commonwealth Fund.

Crombie, I. Davies, HTO (1998). 'Beyond health outcomes: the advantages of measuring process.' *Journal of Evaluation in Clinical Practice*, 4(1): 31–38.

Cutler, T (2002). 'Star or black hole?' *Community Care*, 30 May: 40–41.

Department of Health (2001). *NHS performance ratings: acute trusts 2000/01*. London: Department of Health (http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4003181).

Department of Health (2002). *NHS performance ratings and indicators: acute trusts, specialist trusts, ambulance trusts, mental health trusts 2001/02*. London: Department of Health (http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4002706).

Dolan, P. Gudex, C. Kind, P. Williams, A (1996). 'Valuing health states: a comparison of methods.' *Journal of Health Economics*, 15(2): 209–231.

Dr Foster Intelligence (2007) [web site]. *How healthy is your hospital?* London: Dr Foster Intelligence (http://www.drfosterintelligence.co.uk/library/reports/hospitalGuide2007.pdf).

Freudenberg, M (2003). *Composite indicators of country performance: a critical assessment*. Paris: Organisation for Economic Co-operation and Development (STI Working paper DSTI/DOC 2003/16).

Halasyamani, L. Davis, M (2007). 'Conflicting measures of hospital quality: ratings from 'hospital compare' versus 'best hospitals'.' *Journal of Hospital Medicine*, 2(3): 128–134.

Healthcare Commission (2004). *2004 performance ratings*. London: Healthcare Commission (http://ratings2004.healthcarecommission.org.uk/).

Healthcare Commission (2005). *2005 performance ratings*. London: Healthcare Commission (http://ratings2005.healthcarecommission.org.uk/).

Healthcare Commission (2007). *Annual health check 2006/07. A national overview of the performance of NHS trusts in England*. London: Healthcare Commission (http://www.cqc.org.uk/_db/_documents/Annual_health_check_national_overview_2006–2007.pdf).

Health Consumer Powerhouse & Frontier Centre for Public Policy (2008). *Euro-Canada health consumer index 2008*. Brussels, Ottawa & Winnipeg (FC Policy Series No. 38).

Institute of Medicine (2001). *Crossing the quality chasm: a new health system for the 21st century*. Washington DC: Institute of Medicine of the National Academies, Committee on Quality of Health Care in America.

Jacobs, R. Goddard, M. Smith, PC (2005). 'How robust are hospital ranks based on composite performance measures?' *Medical Care,* 43(12): 1177–1184.

Jencks, S. Cuerdon, T. Burwen, D. Fleming, B. Houck, P. Kussmaul, A. Nilasena, D. Ordin, D. Arday, D (2000). 'Quality of medical care delivered to Medicare beneficiaries: a profile at state and national levels.' *Journal of the American Medical Association,* 284(13): 1670–1676.

Jencks, S. Huff, E. Cuerdon, T (2003). 'Change in the quality of care delivered to Medicare beneficiaries, 1998–1999 to 2000–2001.' *Journal of the American Medical Association,* 289(3): 305–312.

Joint Research Centre (2002). *State-of-the-art report on current methodologies and practices for composite indicator development*. Report prepared by the Applied Statistics Group. Brussels: European Commission, Institute for the Protection and Security of the Citizen (http://composite-indicators.jrc.ec.europa.eu/Document/state-of-the-art_EUR20408.pdf).

Kmietowicz, Z (2003). 'Star rating system fails to reduce variation.' *British Medical Journal*, 327(7408): 184.

Krumholz, H. Rathore, S. Chen, J. Wang, Y. Radford, M (2002). 'Evaluation of a consumer-orientated internet health care report card.' *Journal of the American Medical Association*, 287(10): 1277–1287.

Lied, T. Malsbary, R. Eisenberg, C. Ranck, J (2002). 'Combining HEDIS indicators: a new approach to measuring plan performance.' *Health Care Financing Review,* 23(4): 117–129.

Lindenauer, P. Remus, D. Roman, S. Rothberg, M. Benjamin, E. Ma, A. Bratzler, D (2007). 'Public reporting and pay for performance in hospital quality improvement.' *New England Journal of Medicine,* 356(5): 486–496.

Lun, G. Holzer, D. Tappeiner, G. Tappeiner, U (2006). 'The stability of rankings derived from composite indicators: analysis of the 'IL Sole 24 Ore' quality of life report.' *Social Indicators Research,* 77(2): 307–331.

Miller, N (2002). 'Missing the target.' *Community Care*, 21 November: 36–38.

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

Nardo, M. Saisana, M. Saltelli, A. Tarantola, S. Hoffman, A. Giovanni, E (2005). *Handbook on constructing composite indicators: methodology and user guide*. Paris: Organisation for Economic Co-operation and Development (OECD Statistics Working Paper 2005/03).

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

Nietert, P. Wessell, A. Jenkins, R. Feifer, C. Nemeth, L. Ornstein, S (2007). 'Using a summary measure for multiple quality indicators in primary care: the Summary QUality InDex (SQUID).' *Implementation Science,* 2: 11.

Nord, E (2002). 'Measures of goal attainment and performance: a brief, critical consumer guide.' *Health Policy*, 59(3): 183–191.

O Brien, S. Shahian, D. Delong, E. Normand, SL. Edwards, F. Ferraris, V. Haan, C. Rich, J. Shewan, C. Dokholyan, R. Anderson, R. Peterson, E (2007). 'Quality measurement in adult cardiac surgery: Part 2 – statistical considerations in composite measure scoring and provider rating.' *Annals of Thoracic Surgery,* 83(Suppl. 4): 13–26.

O Malley, A. Zaslavsky, A. Hays, R. Heppner, K. Keller, S. Cleary, P (2005). 'Exploratory factor analyses of the CAHPS® hospital pilot survey responses across and within medical, surgical, and obstetric services. *Health Services Research,* 40(6 Pt 2): 2078–2095.

Page, S. Cramer, K (2001). 'Maclean's rankings of health care indices in Canadian communities, 2000: comparisons and statistical contrivance.' *Canadian Journal of Public Health* (*Revue Canadienne de Sante Publique)*, 92(4): 295–298.

Premier (2005). CMS/ Premier Hospital Quality Incentive Demonstration (HQID). Washington, DC: Premier (http://www.premierinc.com/all/quality/hqi/index.jsp).

Reeves, D. Campbell, S. Adams, J. Shekelle, P. Kontopantelis, E. Roland, M (2007). 'Combining multiple indicators of clinical quality: an evaluation of different analytic approaches.' *Medical Care,* 45(6): 489–496.

Ryan, M. Farrer, S (2000). 'Using conjoint analysis to elicit preferences for health care.' *British Medical Journal*, 320(7248): 1530–1533.

Saaty, R (1987). 'The analytical hierarchy process: what it is and how it is used.' *Mathematical Modelling*, 9: 161–176.

Smith, PC (2002). Developing composite indicators for assessing health system efficiency. In: Smith, PC (ed.). *Measuring up: improving the performance of health systems in OECD countries*. Paris: Organisation for Economic Co-operation and Development.

Snelling, I (2003). 'Do star ratings really reflect hospital performance?' *Journal of Health Organization and Management,* 17(3): 210–223.

US News & World Report (2007). *America's Best Hospitals 2007 methodology.* Washington, DC. New York, NY: US News & World Health Report (http://www.usnews.com/usnews/health/best-hospitals/methodology_report.pdf).

Werner, R. Bradlow, E (2006). 'Relationship between Medicare's hospital compare performance measures and mortality rates.' *Journal of the American Medical Association,* 296(22): 2694–2702.

WHO (2000). *The world health report 2000. Health systems: improving performance.* Geneva: World Health Organization.

Williams, A (2001). 'Science or marketing at WHO? A commentary on world health 2000.' *Health Economics,* 10(2): 93–100.

Williams, S. Koss, R. Morton, D. Loeb, J (2006). 'Performance of top-ranked heart care hospitals on evidence-based process measures.' *Circulation*, 114(6): 558–564.

Zaslavsky, AM. Shaul, JA. Zaborski, LB. Cioffi, MJ. Cleary, PD (2002). 'Combining health plan performance indicators into simpler composite measures.' *Health Care Financing Review,* 23(4): 101–116.