Medical and Information Technology

Intelligent Information: A National System for Monitoring Clinical Performance

Alex Bottle and Paul Aylin

Objective. To use statistical process control charts to monitor in-hospital outcomes at the hospital level for a wide range of procedures and diagnoses.

Data Sources. Routine English hospital admissions data.

Study Design. Retrospective analysis using risk-adjusted log-likelihood cumulative sum (CUSUM) charts, comparing each hospital with the national average and its peers for in-hospital mortality, length of stay, and emergency readmission within 28 days.

Data Collection. Data were derived from the Department of Health administrative hospital admissions database, with monthly uploads from the clearing service.

Principal Findings. The tool is currently being used by nearly 100 hospitals and also a number of primary care trusts responsible for purchasing hospital care. It monitors around 80 percent of admissions and in-hospital deaths. Case-mix adjustment gives values for the area under the receiver operating characteristic curve between 0.60 and 0.86 for mortality, but the values were poorer for readmission.

Conclusions. CUSUMs are a promising management tool for managers and clinicians for driving improvement in hospital performance for a range of outcomes, and interactive presentation via a web-based front end has been well received by users. Our methods act as a focus for intelligently directed clinical audit with the real potential to improve outcomes, but wider availability and prospective monitoring are required to fully assess the method's utility.

There is an ever-increasing focus on monitoring clinical standards in many countries' health services, including measures of process (such as administration of prophylactic antibiotics), outcome (such as mortality), and safety (adverse events and patient satisfaction). Data are commonly taken from hospitals' administrative data sets, from other bodies such as national clinical audits, or collected specifically for the measurement of performance indicators. The

range, type, construction, and use of such indicators vary greatly across countries, although some, such as mortality following acute myocardial infarction, are likely to be common to many westernized nations due to the (relative) “hardness” of the endpoint (death) and of the diagnosis and because there is a sound evidence base regarding treatment protocols.

We describe a performance monitoring tool used by nearly 100 National Health Service (NHS) hospitals in England (Dr. Foster Intelligence 2005). We begin by outlining the main approaches in the United States and the United Kingdom in terms of who sets the clinical standards, what metrics and benchmarks are used to assess performance, and how the feedback loop is closed, before presenting the tool. We first describe the role of the tool and how it fits in with existing U.K. policy, the types of user and some examples of how the tool has been used by hospitals to improve their outcome rates. We then give the technical details covering the data sources, the outcome measures monitored and statistical methods regarding the statistical process control charts including a consideration of case-mix adjustment, and how the results might be acted upon (operationalized) by a hospital using the scheme.

U.S. APPROACH

Setting of Standards

The Joint Commission on Accreditation of Healthcare Organizations is the nation’s predominant standards-setting and accrediting body in health care. Accreditation by the Joint Commission is recognized nationwide as a symbol of quality that indicates that an organization meets certain performance standards. Also important is the National Quality Forum, created in 1999 to “improve American health care through endorsement of consensus-based national standards for measurement and public reporting of health care performance data.” Their “Compendium 2000–2005” covers all their endorsed measures and standards (Quality Forum 2006).

To determine whether health care plans meet these standards, the calculation of measures is required, and the choice and use of these measures come not from the institutions that created them but from government bodies such as the Centers for Medicare and Medicaid Services (CMS) and the purchasers of care.

Address correspondence to Alex Bottle, Ph.D., Dr. Foster Unit at Imperial College, Department of Primary Care and Social Medicine, Imperial College London, First floor, Jarvis House, 12 Smithfield Street, London EC1A 9LA, U.K. Paul Aylin, M.B., Ch.B., F.F.P.H.M., is with the Dr. Foster Unit at Imperial College, Department of Primary Care and Social Medicine, Imperial College London, London, U.K.

Assessment of Meeting Standards

In 1997, the Joint Commission launched “ORYX: The Next Evolution in Accreditation” to integrate the use of outcomes and other performance measures into the accreditation process (Joint Commission on Accreditation of Healthcare Organizations 2006). A component of the ORYX initiative is the identification and use of standardized (“core”) performance measures; accredited hospitals began collecting data on these in 2002. The hospital core measures for 2006 covered acute myocardial infarction, heart failure, pregnancy and related conditions, pneumonia, and surgical infection prevention. To earn and maintain accreditation, an organization must undergo a regular on-site survey by a Joint Commission survey team at least every 3 years. However, hospitals pay for Joint Commission surveys, and more than 70 percent of the Joint Commission’s revenue comes directly from the organizations it is supposed to inspect (American Federation of Teachers 2006). The Joint Commission have switched to unannounced inspections for all hospitals during the 3-year cycle, but these have been criticized for being superficial and failing to detect significant hospital safety and performance problems—identification of poor care patterns are not distinguishable by the Full Survey score or by the accreditation decision (United States Government Accountability Office 2004; Moffett, Morgan, and Ashton 2005).

The Agency for Healthcare Research and Quality (AHRQ) has developed a large number of quality indicators such as mortality rates for stroke and coronary artery bypass graft (CABG) for use with hospital data that can enable hospitals and also federal and state policy makers to track performance over time (AHRQ 2006). These were developed and expanded from the Healthcare Cost and Utilization Project, which aims to build uniform databases from hospital-based administrative data. The Agency also supplies software to calculate the indicators of interest to a given hospital once that hospital has extracted the necessary data.

When large employers purchase health care for their staff, they want value for money, which has led to the National Committee for Quality Assurance (NCQA) overseeing the Health Plan Employer Data and Information Set (HEDIS). This is a tool used by more than 90 percent of America’s health plans to measure performance on important dimensions of care and service (NCQA 2006a). Health care plans are given scores for 17 different performance measures. These include how well the plans manage high blood pressure or how precisely they adhere to clinical evidence-based protocols. The NCQA publish their national benchmarks and national and regional

thresholds for HEDIS measures for each accreditation year (NCQA 2006b). NCQA determines the HEDIS measures' portion of the score by comparing the provider's results with the benchmark of the 90th percentile of national results and with regional and national thresholds (the 75th, 50th, and 25th percentiles). CMS accreditation uses similarly calculated HEDIS benchmarks and thresholds.

The most prominent example of providers themselves measuring performance in order to improve is the nation's largest health care provider, the Department of Veterans Affairs. Beginning in the early 1990s, they established system-wide quality improvement initiatives, many of which the Institute of Medicine would later recommend. An example is their National Surgical Quality Improvement Program, which uses performance measurements, reports, self-assessment tools, site visits, and best practices (American College of Surgeons 2006).

Publication and Completing the Feedback Loop

If hospitals do not meet the prescribed standards, then CMS can withdraw their accreditation and eligibility for federal hospital funding and the Joint Commission may withdraw their accreditation, leading to the loss of private health carrier reimbursements, until corrective changes are made.

A recent important development is the emergence of pay for performance. Some organizations such as CMS have incorporated selected AHRQ indicators into this process. In the Premier Hospital Quality Incentive Demonstration, hospitals scoring in the top 10 percent for quality measures relating to five clinical conditions will receive a 2 percent bonus payment on top of the standard DRG payment. Those scoring in the next highest 10 percent will receive a 1 percent bonus. In the third year of the project, those hospitals that do not meet a predetermined threshold score will see their payments reduced (CMS 2005). Hospital-specific performance will be publicly reported on CMS's website. The project is a demonstration and involves the voluntary participation of over 260 hospitals and is designed to determine if financial incentives are effective at improving the quality of inpatient hospital care.

Patients can access a web-based program called Hospital Compare (Hospital Quality Alliance and the United States Department of Health and Human Services 2006) to see how any given participating hospital compares for heart attack, heart failure, pneumonia, and surgery indicators with the averages for the nation, the state, and the top 10 percent of hospitals. Data are provided by the Hospital Quality Alliance, which encourages hospitals to

collect and publish their data voluntarily. Graphs for the hospital with the three comparisons drawn on are shown, with tables giving the denominators.

U.K. APPROACH

Setting of Standards and Targets

The 24 (with component parts) “core” standards for providers of NHS services, such as “Healthcare organizations must enable all members of the population to access services equally and offer choice in access to services and treatment equitably,” outline the acceptable level of care as set by the Department of Health, who also set the current national targets in 2004 (Department of Health 2004). These comprise standards that all health care organizations in England that treat NHS patients should be achieving now and “developmental standards” that they should be aiming to achieve in the future. It is the responsibility of trust boards to satisfy themselves that they are meeting core standards and, where this is not happening, to take appropriate steps to correct the situation (Healthcare Commission 2005a).

Assessment of Meeting Standards and Targets

In the United Kingdom until 2005, hospitals were compared using the star rating system of the Healthcare Commission, originally the Commission for Healthcare Improvement, an independent body set up to promote and drive improvement in the quality of health care and public health (Healthcare Commission 2006a), charged with assessing the performance of every NHS hospital trust and private health care provider. This system awarded each NHS hospital up to three stars by combining a large number of diverse indicators covering a range of services, varying from the administrative, such as financial management, to the clinical, such as waiting times for referral for suspected cancer. Each hospital was assigned to one of five bands according to the position relative to the national average of three different confidence intervals (CIs) around the rates. Hospital results were available online as a band for each indicator or the overall number of stars (Healthcare Commission 2005b). This has been replaced with the “annual health check,” based upon measuring performance within the Department of Health framework of national standards and targets and intended to be more “patient centered” (Day 2006). Core standards, existing national targets, use of resources and new national targets are scored separately. The responsibility is placed on boards of trusts to make a self-assessment and public declaration on the extent to which

their organization has met the core standards. To measure performance against the 21 existing national targets, the Healthcare Commission used 26 different indicators in the 2005/2006 annual health check, with 13 applicable to acute and specialist hospital trusts, such as whether the patient waited more than 3 months for revascularization or more than 4 hours in accident and emergency.

A number of indicators, such as mortality following heart bypass and emergency readmissions following hip fracture, are derived from routine administrative data (Hospital Episode Statistics [HES]) that all NHS hospitals are mandated to collect and submit at least quarterly to the Department of Health. Our tool also uses these data. The set of indicators is also informed by the “better metrics project,” begun in 2004 to improve the clinical relevance of NHS performance assessment measures, and to date covering 11 clinical areas and health inequalities (Whitty et al. 2006). The annual health check has two elements: quality and use of resources. The quality element sums the scores for the assessment of whether core standards and existing or new national targets are met, using the following four-grade scale: “fully met,” “almost met,” “partly met,” or “not met.” In terms of quality overall, hospital trusts are then rated as “excellent,” “good,” “fair,” or “weak.”

Publication and Completing the Feedback Loop

The Healthcare Commission checks these annual health check self-declarations against a wide range of surveillance information and will follow up if there are discrepancies between the two sources; after the 2005/2006 results, 60 out of a total of 570 NHS trusts were inspected for this reason, with another 60 inspected after being chosen at random. The inspections looked at whether the documentary evidence that the trusts relied upon when making their declarations was adequate. Following inspection, recommendations were made and outcome measures by which the practice changes would be judged were described. Results by health care provider from the annual health check (and its predecessor) are freely available on the Internet.

A recent addition to publicly accessible hospital performance data covers survival following cardiac bypass and aortic valve replacement using data from the Society for Cardiothoracic Surgeons in Great Britain and Ireland (Healthcare Commission 2006b). This gives risk-adjusted survival rates by center and named surgeon, though the data are not easy to extract and the Society admits that, as it used EuroSCORE for risk adjustment, risks are overpredicted due to technical improvements in surgery and anesthetics.

The Society has recently introduced a voluntary accreditation scheme, involving site visits and comparison of risk-adjusted outcome rates against the Society's targets, with mechanisms of dealing with underperformance (Cardiothoracic Surgery Network 2004).

Private Hospitals

Private hospitals (i.e., those not in the NHS) are most commonly used for elective surgery in the United Kingdom. The Healthcare Commission currently has a statutory obligation to inspect all independent sector registered establishments at least annually, and is continuing to develop the process for registering and inspecting. They have recently developed a series of high-level indicators to help monitor the performance, such as overall perioperative mortality and surgical site infections. Some of this information is also submitted to bodies such as the Independent Healthcare Forum (as part of their credentialing program for the private sector) and the U.K. arm of the Quality Indicator Project, which originated in the United States and now operates in 12 countries (U.K. Quality Indicator Project 2006). Our system does not yet cover private hospitals as we do not have the data.

ROLE AND USERS OF OUR TOOL

We view our system as complementary to government-led performance monitoring. As well as being voluntary to join, it does not detail improvements to *processes* that must be made, whereas participation in the regulator's "annual health check," like the star ratings system before it, is mandatory and focuses on processes. Rather, our tool flags particular diagnoses or operations with significantly high (or low) *outcome* rates at the user's unit. Unlike with Hospital Compare and the "annual health check," output from our tool is not published. We have chosen to concentrate on outcome measures as these are generally what matters most to the patient and are easier to measure from routine data. The onus is then on the user to investigate further as described later, beginning with subanalyses using the tool and proceeding to inspection of case notes and processes along the care pathway within the hospital. If managers and clinicians are serious about improving quality, they will not ignore the tool's findings. If quality of care is found to be substandard to the point of leading to undesirable outcomes, then improvement efforts may be assessed over time using the tool.

As well as hospital trusts, the system is also used by primary care trusts and strategic health authorities. Among their primary responsibilities is the assessment of their patients' needs and purchasing hospital care for them. Although there is no competitive market for emergency care in the United Kingdom, hospitals do now effectively compete for contracts for elective work with the private sector and in particular the new Independent Sector Treatment Centres. Strategic health authorities can demand value for money from hospitals that serve their patients, and there is particular focus currently on reducing inappropriately long length of stay, thereby reducing costs. One strategic health authority has asked all its hospital trusts to sign up to our tool for this reason (see "real-world examples of use"). More and more hospitals are seeking to become Foundation Trusts, which gives them, for example, more financial autonomy from the government while remaining within the NHS, but applications to become Foundation Trusts are only accepted from top-performing hospitals. One hospital trust has used benchmarking results from our tool as part of their application. These policies combine to put pressure on hospitals to improve their outcomes.

REAL-WORLD EXAMPLES OF USE

Procedure-Specific Mortality Alarms

The medical director of one hospital trust was alerted by an alarm (when the chart crosses the threshold—see "Tool Methodology" for details on chart construction) on their alarm screen (as in Figure 2) for high in-hospital mortality for lower gastrointestinal surgery. On further analysis using the tool's drill-down (cross-tabulation) capability, it was clear that emergency admissions occurring on some days of the week had a significantly higher risk of death than on others. A review of the rota system revealed that the most experienced surgeons were not available at these times for operating on the most severe cases. The rota was changed to allow the most experienced surgeons to be free to cover for these days. Mortality has since dropped to average levels, with no further chart alarms.

Disease-Specific Readmission Alarms

The Director of Performance and Development at another trust followed up an alarm regarding readmissions for chronic obstructive pulmonary disease. Activity was reviewed for a 9-month period. There were 87 readmissions

against an expectation of 71. A review of the notes identified that six patients accounted for 31 admissions and that there were significant clinical factors associated with each case. It was concluded that these patients did indeed have severe disease that warranted the extra hospitalizations and that no further action was required.

Mortality Alarms Involving Multisite Hospitals

Two hospitals with high overall mortality demonstrated by the tool decided on more far-reaching action. Walsall Hospitals NHS Trust's medical director and others formed seven clinical governance groups to implement changes in several clinical disease areas. Similarly, changes were initiated in several management areas including audit department, clinical risk, continuing professional development unit, bed management, and information services, and the significant decrease in mortality can be seen on the tool (Jarman et al. 2005). Bradford Teaching Hospitals NHS Trust established a hospital mortality reduction group with senior leadership. The tool was used with death certificates and local routine hospital data to review progress. There was extra training in areas such as clinical observation, medication safety, and infection control (Wright et al. 2006).

Length of Stay Alarms Involving Multiple Hospitals

Another type of user is the Strategic Health Authority, of which there are currently 10 in England, with responsibility for clinical governance and in getting the best care for patients within its boundaries. One such Authority asked all its hospital trusts to sign up for the tool and found that many of its acute trusts had a lot of alarms for length of stay. As they drilled down, they found that short lengths of stay were associated with better clinical outcomes. After retrieving case notes and checking the data, long lengths of stay in some cases were due to delays in accessing investigations and, at its worst, sometimes led to increased mortality. One example of this was the treatment of fractured neck of femur. The tool enabled them to analyze mortality and length of stay by preoperative length of stay, a measure which, if prolonged, has been found to be associated with higher mortality (Bottle and Aylin 2006). Some of the Authority's acute trusts had average lengths of stay of up to 36 days, but in 6 months that particular trust reduced its average from 36 to 18 days by focusing on the clinical pathway.

USE OF THE TOOL IN PRACTICE: FOLLOWING UP A SUSPECTED HIGH OUTCOME RATE

We now describe how the user, usually hospital managers but also senior clinicians, accesses the system, requests the desired analyses, and then “operationalizes” the results by investigating a chart alarm and completing the loop.

When the user gets past the log-in screen after typing in their user name and password supplied by the system administrator, they see a grid with diagnosis and procedure groups as rows and outcome measures as columns, with each cell containing either a red bell symbol if the chart has crossed the threshold for an odds ratio (OR) of 2 in the last 3 months, suggesting a high outcome rate, or a green bell symbol if the chart has crossed the threshold for an OR of 0.5, suggesting a low outcome rate. If the threshold has not been crossed for a given diagnosis or procedure, the cell is not shown (Figure 1). Clicking on a bell will display the relevant chart (see “Tool Methodology”) with an accompanying table of figures giving observed and expected outcomes with CIs for the monitored period. Charts are plotted chronologically

Figure 1: Screenshot of Online Monitoring System Front Page That the User Sees after Logging In

and will come to a stop at the most recent patient, whose date of discharge may be read from the chart.

Each alarm is a starting point for action. In the ideal scenario, all alarms would represent genuine problems with performance, but in practice some will be false alarms. These could be considered to be of two types. "Statistical" false alarms, in which the trust's outcome rate is in fact compatible with the national average but the trust has had a run of "bad luck," i.e. the alarm occurred by chance; these can be minimized by raising the chart threshold. Further alarms after resetting the chart rather than a single alarm would suggest a genuinely high rate. "Medical" false alarms occur when the odds of the outcome of interest are at least twice the benchmark but the cause is not one of poor quality of care.

In real practice, however, we do not know if an alarm is false (what in screening terms would be called a false positive) or the detection of an outcome with at least twice the benchmark odds (true positive, though some of these will be medical false alarms).

The following has been suggested as a "check list" for the hospital leaders following up an alarm (Marshall and Mohammed 2002; Lilford et al. 2004):

- (1) Check data quality.
- (2) Assess the case mix.
- (3) Consider policy or organizational ("process of care") issues ("structure" in Lilford and colleagues's pyramid).
- (4) Quality of care.

After an alarm, the probability of a false alarm given the number of patients monitored and underlying outcome rate can be displayed from tables of results from prerun simulations. If this is felt to be high, then the chart may be reset with no further action. If not, or if the chart again crosses the threshold soon after it first does so, then the data quality should be checked, for example using the data quality report that is part of our tool (e.g., to see if a high proportion of records have been excluded due to invalid or duplicate entries) and then by comparing the admissions and outcomes on the tool with first the hospital's electronic records and then the patient notes.

If the first stage does not reveal an explanation for the alarm, the second step is to examine the case mix of the patients, for example by auditing a random sample. Clinicians may be aware of other case-mix issues such as specializing in palliative care. At this stage, organizational issues can be considered such as the appropriateness of referral and delays in admission beyond

the control of the hospital. If these can be excluded or are found to be of insufficient magnitude to explain the high outcome rate, then quality of care could be the explanation, from inappropriate referral and preoperative care through to peri- and postoperative care. Drop-down menus guide the user to selecting the time period, age range, weekday of admission, and many other factors to see if the alarm affects all patients with the diagnosis or procedure or just a subgroup. The time period of monitoring can be extended back to any time from 1996 in order to see if the trust has had an earlier alarm and how many it had to see how long-standing the potential problem is. Data can be viewed by individual named consultant (depending on the access privileges set by the administrator) and compared with others in the same specialty.

Although the national average is used for all inpatient benchmarks, each trust can compare itself with six of its peers, defined as having either similar volume by the user themselves, or the six best performers in a table giving the relative risks for each trust.

After looking at aggregate admission counts and observed and expected outcomes, the authorized user (with access to specific patient data) can view the complete electronic record for each admission, including all diagnoses and procedures, dates, age, sex, and outcomes, which can be downloaded into Excel. This can help with validation of the electronic data against the patient notes (item 1 on the above check list) and a starting point for considering case-mix issues (item 2) and then the care pathway (items 3 and 4).

TOOL METHODOLOGY

Data Sources

HES are routinely collected data that cover all inpatient and day case admissions to NHS hospitals in England. The monitoring tool uses 9 years of admissions data from 1996/1997 to 2004/2005, covering acute and community hospital trusts (a trust can consist of several hospitals, each with their own site code). These are augmented by monthly submissions from each trust via the NHS-wide Clearing Service (a data warehouse) from April 2005 so that data are at most 6 weeks out of date at any time.

Each of the 14 diagnosis fields is coded using ICD10 and we assign it to one of 259 clinically meaningful groupings using the AHRQ's CCS classification (Agency for Health Care Policy and Research 2003). The 12 operation fields use U.K. OPCS4 codes (Office of Population Censuses and Surveys

1990), of which the first is usually the most major even if it was not the first to be performed. No grouping scheme for OPCS4 codes currently exists, and we have therefore grouped them together after taking clinical advice from a number of professional bodies (e.g., the Vascular Surgical Society). Not all diagnosis and procedure groups have enough numbers to enable robust comparisons and the tool monitors 77 diagnosis and 102 procedure groups. Groups were usually chosen for monitoring if they had large numbers of deaths or admissions, but a few less common procedures were requested by clinicians. The diagnosis groups cover over 80 percent of deaths and admissions, and the procedure groups about 70 percent of deaths and 80 percent of admissions with some procedure recorded.

The basic unit of the database is the consultant episode, the continuous period of time during which the patient is under the care of a particular consultant, whose registration number is recorded, enabling consultant-level analysis. In an admission, a patient can have any number of episodes, though around 85 percent have only one. Episodes are linked together into admissions if they belong to the same patient and have the same admission and discharge date at the same trust, and admissions are linked together to account for interhospital transfers.

The diagnosis used for monitoring is the first field ("primary diagnosis") for the first episode, i.e., on admission, although if there is only a vague symptoms and signs diagnosis in this episode, a diagnosis is taken from the subsequent episode (if there is one). All outcome measures are assigned to this diagnosis as the reason for admission is usually of most interest. The procedure used for monitoring is usually the first nonmissing procedure field containing one of the codes in Table 1, with some extra rules concerning cardiac procedures, e.g., CABG takes priority over cardiac catheterization.

Availability and Quality of Data. Ideally, the statistical process control charts could be constructed in real time from patient administration system data so that any necessary remedial action may be taken as early as possible. During the current financial year, hospitals are able to resubmit data, if, for example, they have updated their diagnostic coding. There is considerable variation between hospitals in the frequency of data submission to the clearing house and in the quality of the most recent submission. Clearly, trusts that are able to submit good quality data on time will detect potential problems earlier than others.

One of the features of our tool is the display of basic data quality measures, such as counts of admissions and percentages with a primary

Table 1: Case-Mix Variables Used in the Tool

<i>Variable</i>	<i>Grouping Method or Distinct Values</i>	<i>Comments</i>
Age	5-year bands	Under-1s and those aged 1–4 comprised their own bands
Sex	Male, female (other values excluded)	
Method of admission	Elective, nonelective (emergency, transfer from hospital, maternity event)	
Area-level deprivation	Quintile, with an equal total population in each	Index of Multiple Deprivation 2004
Primary diagnosis	Three- or four-character code	Used in, e.g., AAA (to detect presence of rupture), abdominal hysterectomy (for malignancy), abdominal GI surgery (for malignancy, Crohn's disease and ulcerative colitis)
Financial year		
Month of admission		For respiratory diagnosis groups (easily derived from date of admission)
Palliative care specialty	1 if treatment specialty in any episode in the admission coded to palliative care, 0 otherwise	
Charlson index of comorbidity	Fitted as a factor, capped at 6	
Number of emergency admissions in previous year	Fitted as a factor, capped at 3	Requires linking of admissions to the same patient

AAA, Abdominal Aortic Aneurysm.

diagnosis of R69X (other causes of morbidity and mortality not elsewhere classified). HES data have improved considerably in recent years and we believe that they are of great value in performance monitoring if their drawbacks are taken into account (McKee, Coles and James 1999; Hansell et al. 2001). Past experience shows that the quality of data improves with use.

Outcome Measures

For diagnosis groups we use death, length of stay, and emergency readmission to any hospital within 28 days of discharge from the final posttransfer hospital as outcomes. For the Audit Commission “basket” of procedures that should

mainly be performed as day-case surgery (Audit Commission 2001), the outcome is the procedure being performed as an inpatient. For other procedure groups we use death within 30 days of the procedure, length of stay, and emergency readmission as per the diagnoses. These outcomes are available from the data and are known to be important. For simplicity, length of stay is dichotomized into whether or not it exceeds the upper quartile for all patients nationally, due to the various problems inherent in trying to treat length of stay as a continuous variable (Yau, Leeb, and Ng 2003). Stays longer than this admittedly arbitrary cut-off point are deemed to be “long” but are common enough to enable robust risk estimation.

Use of Statistical Process Control Charts

Instead of aggregating patient outcomes into annual summaries and comparing each hospital’s outcome rate with the “expected” rate based on the benchmark (an “acceptable” level of performance, often simply the national average), individual-level control charts plot patient by patient a function of the difference between their actual outcome and their expected (a priori) probability or risk of having that outcome. These charts are run for as long as is desired in order to have sufficient power to detect a difference between the observed and expected outcome rates. There are a variety of charts available, but the log-likelihood cumulative sum (CUSUM) is the most powerful test for detecting unacceptably high rates for a given false-positive rate (Moustakides 1986). For this and other reasons discussed elsewhere (Marshall et al. 2004) we have adopted the log-likelihood method of Steiner et al. (2000), which includes adjustment for the a priori risk according to whatever case-mix variables are available. This chart requires that the following issues be considered, which are now discussed: estimation of the expected including case-mix adjustment, setting of benchmarks and the chart threshold, and what to do when the threshold is crossed.

Estimation of the Expected Risk of Each Outcome

For each diagnosis or procedure group and outcome, logistic regression models were constructed using the data for 1996/1997 to 2004/2005. Case-mix information in HES is limited because the data set was created to measure activity, but age, sex, method of admission (whether emergency or elective), quintile of a multiple socioeconomic deprivation score, IMD2004, for the area (electoral ward) of residence (Office of the Deputy Prime Minister Indices of Deprivation 2004) and diagnosis or procedure subgroup (e.g., asthma was

divided into asthma and status asthmaticus) were available and were entered into the model (Table 1). For the respiratory groups, the month of admission was also included. All patients within the same risk stratum (combination of age, sex, etc.) were therefore allocated the same risk. The success of case-mix adjustment for accurately predicting the outcome (discrimination) was evaluated using the area under the receiver operating characteristic curve (c statistic). This is between 0.5 (discrimination being no better than chance) and 1 (perfect discrimination); values below 0.7 are considered poor or fair, 0.7–0.8 considered good, and higher values considered very good or excellent.

The area under the curve (c) statistics were generally good or very good for mortality, but only fair for emergency readmission. In a study comparing mortality prediction using administrative and clinical data sets, correlations between hospital-level observed-to-expected ratios ranged for noncardiac surgery from 0.64 to 0.86 depending on the specialty (Gordon et al. 2005). Compared with the clinical data sets, the administrative ones identified outlier hospitals with a sensitivity of 73 percent and a specificity of 89 percent. For CABG surgery, Geraci et al. (2005) found that administrative data gave a c statistic of 0.70 compared with 0.76 for the clinical data set, but that adding just two variables (previous heart surgery and whether the surgery was elective, urgent or emergency—in English data we distinguish between elective and emergency admission) increased the c statistic to 0.74. More sophisticated risk scoring systems have encountered significant problems such as high complexity and overpredicting risk, and there is some evidence that simple methods may suffice (Sutton et al. 2002).

Benchmarks and Setting the Chart Threshold

In the absence of agreed benchmarks for our outcome measures, we compare each trust with the national average. The charts aim to detect twice or over the national odds for poor performance and half or under the national odds for good performance. If the patient dies, the chart moves up by an amount inversely proportional to their a priori risk of death so that the trust is not unduly penalized when very ill patients die, and moves down if they survive. The higher the chart rises, the more likely it is that the odds of death for the trust are twice the national average: a lower threshold gives speedier detection of high odds but at the cost of a higher false alarm rate. This trade-off between successful detection and false alarms was assessed by simulation; the emphasis was on suppressing the false alarm rate because a large number of false alarms would erode the user's confidence in the tool. There are other measures of

assessing the statistical performance of the chart beyond the scope of this article (see, e.g., Frisén 1992).

When the threshold (“*h*”) is crossed, it is immediately reset to a value of *h*/2, akin to putting the hospital “on probation,” so that if the hospital’s odds continued thereafter to be at least twice the national average, this would be detected more quickly than if the chart had been reset to zero (which would be akin to “wiping the slate clean”). This resetting to *h*/2 has some theoretical justification (Lucas and Crosier 1982). An example chart is given in Figure 2, which shows death from CABG in a sample NHS trust. The OR for the whole period is 1.44 (95 percent CI 0.95–2.09), not significant at the 5 percent level, but without the chart, one period of much higher mortality (when the threshold is crossed) would be concealed.

Further Work

Further improvements to the case-mix model are being evaluated including an exploration of adjustment for comorbidity using the Charlson index (Sundararajan et al. 2004). The introduction of payment by Healthcare Resource

Figure 2: Screenshot of Web Front End Showing the Cumulative Sum (CUSUM) Chart for One Hospital’s Mortality Following Admission for Coronary Artery Bypass Graft (CABG)

Groups (similar to Diagnosis Related Groups, which use all the diagnosis fields and not just the primary diagnosis) provides a financial incentive to encourage the recording of comorbidities, as occurred in the United States (Carter, Newhouse, and Relles 1990). Also potentially of use are the previous admissions or surgery within a given time period, which requires the linking of admissions to the same patient.

We will continue to seek clinical advice in the definitions of procedure groups and development of relevant outcome measures. Recent English admissions data also include Intensive Therapy/High Dependency Unit data, which could be used as a “near miss” outcome instead of or in combination with death (Steiner, Cook, and Farewell 1999). We dichotomized length of stay so that an “event” was a stay of more than the upper quartile length of stay for all patients in order to simplify the analysis. However, such categorization loses information of course, and, although there are very different approaches to modeling it as a continuous variable (e.g., Marazzi et al. 1998; Wang, Yau, and Lee 2002), further work could lead to a fuller understanding of a hospital’s length of stay distribution.

Although monetary expenditure and clinical performance are very different in nature, the web-based drill-down flexibility would also suit analyses of financial flows, and we have a related tool in development that uses current Healthcare Resource Group tariffs to track financial flows.

In light of a government initiative within the United Kingdom in 2004 to offer patients elective care appointments from a choice of five hospitals (Department of Health 2004), we are now working to provide Internet-based summary analyses using key indicators. This will assist both the patient and general practitioner in choosing their hospital of treatment.

CONCLUSIONS

We have created a system that allows the monitoring of clinical outcomes with a short time lag, with considerable advantages over more traditional league tables that are still sometimes used in the United Kingdom or performance rating systems used previously in health care in England. This system allows for:

- Analysis of timely data, updated monthly rather than annually.
- Use of the most statistically powerful tests for successful, automated detection of problems at the earliest opportunity (including a quantifiable screening process for false alarms).

- Interactive front-end capabilities with drill-down options that allow for enhanced clinical decision making that directly impacts on quality of patient care and hospital levels of performance.

We envisage the system as a management tool for clinicians and managers, offering prospective near real-time monitoring of different outcomes within hospitals. It could act as a focus for intelligently directed clinical audit with the real potential to reveal both problems and good practice well in advance of the U.K. Healthcare Commission's "annual health check" or similar governmental assessments. The usability of the tool's front end is important so that the retrieval of well-presented key information is as quick and easy as possible. There is some evidence from hospitals such as that given earlier suggesting that the analyses available within the tool relate to clinical experience and enhance decision making, but wider availability and prospective monitoring will be required to fully assess the utility and impact on clinical practice. Health information technology is increasingly being considered by state leaders in the United States too to improve health care via public-private collaboration (Virtual Medical Worlds Monthly 2006).

ACKNOWLEDGMENTS

We are grateful to Joanne Zaborowski at the Capital Health Center in Edmonton, Canada for her very helpful review of the manuscript and suggestions for improvement.

Disclosures: The Unit is funded by a grant from Dr. Foster Intelligence (an independent health service research organization).

Disclaimers: None.

REFERENCES

- Agency for Health Care Policy and Research. 2003. "Clinical Classifications Software (ICD-10) Summary and Download. Summary and Downloading Information" [accessed August 2006]. Available at <http://www.ahrq.gov/data/hcup/ccsacd10.htm>
- Agency for Healthcare Research and Quality (AHRQ). 2006. "Quality Indicators" [accessed August 2006]. Available at <http://www.qualityindicators.ahrq.gov/>
- American College of Surgeons. 2006. "National Surgical Quality Improvement Program" [accessed August 2006]. Available at <https://acsnsqip.org/login/default.aspx>

- American Federation of Teachers. 2006. "Joint Commission on Accreditation of Healthcare Organizations" [accessed August 2006]. Available at <http://www.aft.org/healthcare/jcaho/index.htm>
- Audit Commission. 2001. "2001 Day Surgery: Review of National Findings" [accessed April 2007]. Available at <http://www.audit-commission.gov.uk/reports/AC-REPORT.asp?Ca+ID=english%5EHEALTH&prodID=A9E075AF-7BCC-4529-BA78-F0D2F22034BC/>
- Bottle, A., and P. Aylin. 2006. "Mortality Associated with Delay in Operation after Hip Fracture: Observational Study." *British Medical Journal* 332: 947–51.
- Cardiothoracic Surgery Network, The. 2004. "Society of Cardiothoracic Surgeons of Great Britain and Ireland" [accessed August 2006]. Available at <http://www.ctsnet.org/sections/newsandviews/inmyopinion/articles/article-27.html>
- Carter, G. M., J. P. Newhouse, and D. A. Relles. 1990. "How Much Change in the Case Mix Index is DRG Creep?" *Journal of Health Economics* 9 (4): 411–28.
- Centers for Medicare and Medicaid Services (CMS). 2005. "Premier Hospital Quality Incentive Demonstration" [accessed August 2006]. Available at <http://www.cms.hhs.gov/apps/media/press/release.asp?Counter=1343>
- Day, M. 2006. "Three in Five NHS Trusts in England Fail on Basic Care." *British Medical Journal* 333: 114.
- Department of Health. 2004. "Choose & Book—Patient's Choice of Hospital and Booked Appointment" [accessed January 2006]. Available at <http://www.dh.gov.uk/assetRoot/04/08/83/52/04088352.pdf>
- Dr. Foster Intelligence. 2005. *Data Analysis Tools: Real Time Monitoring*. London: Dr. Foster Intelligence.
- Frisén, M. 1992. "Evaluations of Methods for Statistical Surveillance." *Statistics in Medicine* 11: 1489–502.
- Geraci, J. M., M. L. Johnson, H. S. Gordon, N. J. Petersen, A. L. Shroyer, F. L. Grover, and N. P. Wray. 2005. "Mortality after Cardiac Bypass Surgery: Prediction from Administrative versus Clinical Data." *Medical Care* 43: 149–58.
- Gordon, H. S., M. L. Johnson, N. P. Wray, N. J. Petersen, W. G. Henderson, S. F. Khuri, and J. M. Geraci. 2005. "Mortality after Noncardiac Surgery: Prediction from Administrative versus Clinical Data." *Medical Care* 43: 159–67.
- Hansell, A., A. Bottle, L. Shurlock, and P. Aylin. 2001. "Accessing and Using Hospital Activity Data." *Journal of Public Health Medicine* 21 (3): 51–6.
- Healthcare Commission. 2005a. "Assessment for Improvement. The Annual Health Check" [accessed August 2006]. Available at http://www.healthcarecommission.org.uk/_db/_documents/04017226.pdf
- . 2005b. "2005 Performance Ratings" [accessed August 2006]. Available at <http://ratings2005.healthcarecommission.org.uk/>
- . 2006a. "Inspecting Informing Improving: About the Healthcare Commission" [accessed August 2006]. Available at <http://www.healthcarecommission.org.uk/aboutus.cfm>
- . 2006b. "Heart Surgery in Great Britain" [accessed August 2006]. Available at <http://heartsurgery.healthcarecommission.org.uk/>

- Hospital Quality Alliance and the United States Department of Health and Human Services 2006. "Hospital Compare" [accessed August 2006]. Available at <http://www.hospitalcompare.hhs.gov/>
- Jarman, B., A. Bottle, P. Aylin, and M. Browne. 2005. "Monitoring Changes in Hospital Standardised Mortality Ratios." *British Medical Journal* 330: 329.
- Joint Commission on Accreditation of Healthcare Organizations. 2006. "Facts about ORYX: The Next Evolution in Accreditation" [accessed August 2006]. Available at http://www.jointcommission.org/AccreditationPrograms/Hospitals/ORYX/oryx_next_evolution.htm
- Lilford, R., M. A. Mohammed, D. Spiegelhalter, and R. Thomson. 2004. "Use and Misuse of Process and Outcome Data in Managing Performance of Acute Medical Care: Avoiding Institutional Stigma." *Lancet* 363 (9415): 1147–54.
- Lucas, J. M., and R. B. Crosier. 1982. "Fast Initial Response for CUSUM Schemes: Give Your CUSUM a Head Start." *Technometrics* 24 (3): 199–205.
- Marazzi, A., F. Paccaud, C. Ruffieux, and C. Beguin. 1998. "Fitting the Distributions of Length of Stay by Parametric Models." *Medical Care* 36 (6): 915–27.
- Marshall, E. C., N. G. Best, A. Bottle, and P. Aylin. 2004. "Statistical Issues in the Prospective Monitoring of Health Outcomes at Multiple Units." *Journal of the Royal Statistical Society A* 167 (3): 541–9.
- Marshall, T., and M. A. Mohammed. 2002. "Differences in Clinical Performance." *British Journal of Surgery* 89 (8): 948–9.
- McKee, M., J. Coles, and P. James. 1999. "'Failure to Rescue' as a Measure of Quality of Hospital Care: The Limitations of Secondary Diagnosis Coding in English Hospital Data." *Journal of Public Health Medicine* 21 (4): 453–85.
- Moffett, M. L., R. O. Morgan, and C. M. Ashton. 2005. "Strategic Opportunities in the Oversight of the U.S. Hospital Accreditation System." *Health Policy* 75: 109–15.
- Moustakides, G. V. 1986. "Optimal Stopping Times for Detecting Changes in Distributions." *Annals of Statistics* 14: 1379–87.
- National Committee for Quality Assurance (NCQA). 2006a. "The Health Plan Employer Data and Information Set" [accessed August 2006]. Available at <http://www.ncqa.org/programs/hedis/>
- . 2006b. "National Committee for Quality Assurance's Quality Compass" [accessed August 2006]. Available at <http://www.ncqa.org/Info/QualityCompass/index.htm>
- Office of Population Censuses and Surveys. 1990. *Tabular List of the Classification of Surgical Operations and Procedures, Fourth Revision*. London: Stationery Office.
- Office of the Deputy Prime Minister. 2004. "Indices of Deprivation 2004" [accessed April 2007]. Available at <http://www.communities.gov.uk/index.asp?id=1128440>
- Quality Forum. 2006. "Compendium 2000–2005" [accessed August 2006]. Available at <http://www.qualityforum.org/txCompendiumwebforpublic.pdf>
- Steiner, S. H., R. J. Cook, and V. T. Farewell. 1999. "Monitoring Paired Binary Surgical Outcomes Using Cumulative Sum Charts." *Statistics in Medicine* 18: 69–86.
- Steiner, S. H., R. J. Cook, V. T. Farewell, and T. Treasure. 2000. "Monitoring Surgical Performance Using Risk-Adjusted Cumulative Sum Charts." *Biostatistics* 1 (4): 441–52.

- Sundararajan, V., T. Henderson, C. Perry, A. Muggivan, H. Quan, and W. A. Ghali. 2004. "New ICD-10 Version of the Charlson Comorbidity Index Predicted In-Hospital Mortality." *Journal of Clinical Epidemiology* 57: 1288–94.
- Sutton, R., S. Bann, M. Brooks, and S. Sarin. 2002. "The Surgical Risk Scale as an Improved Tool for Risk-Adjusted Analysis in Comparative Surgical Audit." *British Journal of Surgery* 89: 763–8.
- U.K. Quality Indicator Project. 2006. [accessed August 2006]. Available at <http://www.ncl.ac.uk/qip/index.htm>
- United State Government Accountability Office. 2004. "Medicare: CMS Needs Additional Authority to Adequately Oversee Patient Safety in Hospitals. GAO-04-850" [accessed April 2007]. Available at <http://www.gao.gov/new.items/d04850.pdf>
- Virtual Medical Worlds Monthly. "eHI Survey" [accessed August 2006]. Available at <http://www.hoise.com/vmw/06/articles/vmw/LV-VM-08-06-15.html>
- Wang, K., K. K. W. Yau, and A. H. Lee. 2002. "A Hierarchical Poisson Mixture Regression Model to Analyse Maternity Length of Hospital Stay." *Statistics in Medicine* 21: 3639–54.
- Whitty, P., M. Richards, R. Boyle, S. Roberts, G. Alberti, and I. Philp. 2006. "Better Metrics Version 7" [accessed August 2006]. Available at <http://www.healthcarecommission.org.uk/serviceproviderinformation/bettermetrics/suggestedmetrics.cfm>
- Wright, J., B. Dugdale, I. Hammond, B. Jarman, M. Neary, D. Newton, C. Patterson, L. Russon, P. Stanley, R. Stephens, and E. Warren. 2006. "Learning from Death: A Hospital Mortality Reduction Programme." *Journal of the Royal Society of Medicine* 99 (6): 303–8.
- Yau, K. W., A. H. Leeb, and A. S. K. Ng. 2003. "Finite Mixture Regression Model with Random Effects: Application to Neonatal Hospital Length of Stay." *Computational Statistics and Data Analysis* 41 (3): 359–66.