A national clinical indicator database: Issues of reliability and validity

JOANNE L BOOTH AND BRIAN T COLLOPY

Joanne Booth was formerly the Program Manager and Brian Collopy is the Clinical Director of the Australian Council on Healthcare Standards Care Evaluation Program.

Abstract

The introduction of performance (clinical) indicators into the accreditation process by the Australian Council on Healthcare Standards is in keeping with global trends and has enabled the establishment of a National Aggregate Database reflecting standards of care in acute health care organisations. The database contains both quantitative and qualitative information on the processes and outcomes of patient care and changes in practice induced through indicator monitoring. Of fundamental importance to the integrity of the database are the issues of indicator validity, responsiveness and reliability. This paper considers these issues, drawing parallels, as appropriate, to other performance indicator programs and studies.

Introduction

In health care today the challenge is to provide services effectively and efficiently whilst maintaining and also improving the quality of care. Performance (clinical) indicators linked to patient management and outcomes are one means by which quality and effectiveness can be measured and monitored. As described by O'Leary (1995), President, Joint Commission on Accreditation of Health Care Organisations, 'Measurement is the first step towards the derivation of useful performance information that can be used to drive the improvement process.'

In conjunction with medical colleges, the Australian Council on Healthcare Standards (ACHS), through its Care Evaluation Program, has been developing clinical performance measures for use within the ACHS Evaluation and Quality Improvement Program, which are directed at evaluating patient care within acute health care organisations. This is in keeping with similar development programs in the United States, Canada and Malaysia. The establishment and use of

objective performance measures, particularly those directed at the outcomes of patient care, lead to consideration of their use not only as internal quality tools but also as comparative measures of performance between organisations. But, for comparative data to be useful for all stakeholders, it is important to know that the measuring tools are valid, reliable and reproducible and that the data can be translated into effective decision-making and improved patient outcomes (Audet & Denman Scott 1993).

The ACHS clinical indicator program has been challenged in this way. This paper will address this challenge by looking at the fundamental issues of validity, responsiveness and reliability, based on information from the Care Evaluation Program National Aggregate Database, whilst drawing parallels to other indicator programs and studies.

The National Aggregate Database

An essential feature of the clinical indicator program is the operation of a National Aggregate Database. Health care organisations participating in the Evaluation and Quality Improvement Program (the new ACHS accreditation program) submit both quantitative and qualitative indicator data for those indicator sets that are relevant to their services for inclusion into the database. The Paradox-based system collates and analyses numerical values for all indicators as well as information relating to indicator usefulness, relevance and further quality activities. There is, however, a caution. In the world of performance measurement, many databases are being developed and much data are being demanded without a clear understanding of the purpose or content of the database; a case of 'let's just collect everything and we'll work it all out in the morning!' (O'Leary 1995). This approach is clearly one to be avoided.

The ACHS is therefore aiming to have a comprehensive national database which provides to the Australian health care industry meaningful, reliable performance information focused on key areas of clinical practice.

In its first three years of operation, 1993–1995, Hospital Wide Medical Indicator (HWMI) data have been received from 430 Australian health care organisations. As shown in Table 1, this represented nearly 40 per cent of all acute care hospitals and 78 per cent of hospitals with more than 100 beds. In 1995, obstetrics and gynaecology data from 100 hospitals were also included in the database for the first time, providing the largest data set on the processes and outcomes of obstetrics and gynaecology practice ever collated in Australia. With the introduction of more discipline-specific sets of indicators and six-monthly reporting of indicator data for all organisations participating in the Evaluation

and Quality Improvement Program, the size and scope of the database will expand significantly.

Table 1: Number of Australian hospitals and number of those submitting HWMI data, 1993–1995

Number of beds	Australian hospitals	Hospitals on Care Evaluation Program database	Percentage of Australian hospitals on database
1–50	672	135	20.1
51–100	195	127	65.1
>100	214	168	78.5
Total	**1081**	**430**	**39.8**

For the 1996 year, data were received and entered for seven sets of indicators which have all been developed by their respective colleges and committees, in the disciplines of psychiatry, internal medicine, emergency medicine, anaesthetics, day procedures, and obstetrics and gynaecology, these being in addition to the HWMIs. A further five sets in surgery, ophthalmology, rehabilitation medicine, paediatrics and radiology were introduced in 1997, and more sets are still in development. The program thus provides a vast array of monitoring tools to assist clinicians and organisations to evaluate, demonstrate and improve performance. However, for information to be used in this way, it is necessary to identify and develop the best measures, which should, as far as possible, be valid, reliable and useable as part of a meaningful evaluation process.

Indicator validity

The concepts of validity and reliability are, at their simplest level, a concern about the consistency and accuracy of measurement. It is important to know whether an indicator measures what is intended, whether it identifies events that warrant further review, and whether it accurately identifies indicator occurrences from at risk populations. Indicators are, nonetheless, dealing with clinical events and the best clinical opinion is subject to observer variation. This is exemplified by the British epidemiologists, Professors Rose and Barker (1986), who state: 'measurements of disease in life, whether clinical or epidemiological, are often incapable of full validation.' In practice, therefore, validity may need to be assessed indirectly.

An important consideration is that of face validity. The ACHS medical clinical indicators are provider-developed – it is clinical experts, nominated by their relevant colleges, who assure the medical soundness and relevance of the indicators, through a process of extensive consultation and review.

Validity is also considered with regard to content, that is, the extent to which clinical indicators actually reflect the issue of concern. For example, that an occurrence of post-operative pulmonary embolism suggests inadequacy in pre-operative prophylaxis regimes in patients having major surgery. To illustrate this further, in a recent comparative study in New South Wales teaching hospitals, a review of post-surgical wound infection cases was undertaken for 423 patients in four surgical categories, using wound classifications according to both the ACHS Care Evaluation Program definitions (three levels) and the National Nosocomial Infection Surveillance (NNIS) classification system (four levels) (Keogh 1996). Whilst the data were collected according to differing definitions, the results as seen in Table 2 are remarkably similar and the differences considered 'not significant'. This study provides weight to the argument that the Care Evaluation Program nosocomial infection indicators are measuring what is intended.

Table 2: Comparison of post-surgical wound infection data collection using ACHS and NNIS guidelines

	ACHS	NNIS
Numerator	11	13
Denominator	423	423
Rate (%)	2.6	3.07

Content validity is also promoted through the development and use of clearly defined data elements, which allow for the specificity and sensitivity of the information to be maximised. The ACHS requires that data be collected according to given definitions for inclusion in the database. Variations are identified and recorded in prepared results booklets and a Care Evaluation Program data officer meticulously screens all data for inconsistencies in definitions used, following up directly with hospitals if required. This process is supported by ongoing education programs in indicator data collection and use.

In support of this process, during 1995 the Care Evaluation Program excluded 10 per cent of HWMI data received, due to data inconsistencies or variations in the definitions used. Up to 30 per cent of post-operative wound infection data

were also excluded, due to variations relating to wound classification and type of surgery. In most instances this variation was a deliberate choice made by the organisation and identified as such.

A further example of the validation process is provided in a study of the 28-day interval between discharge and readmission for emergency readmissions (Sibbritt 1995). This study, conducted by the Health Services Research Group in New South Wales, related to the HWMI for unplanned readmissions (ACHS 1995). The researchers concluded that, in examining the patterns of readmission to New South Wales public acute hospitals, there was statistical support for use of a 28-day time interval for readmission across four different patient categories of medical, surgical, paediatric, and obstetrics and gynaecology.

Content validity is also determined by extensive field testing of the indicator sets on a national basis before their release. Through the process of testing, feedback and review by college working parties, indicators are modified and/or deleted to ensure that they are measuring what is intended. This field testing process is akin to that of alpha testing of indicators by the Joint Commission on Accreditation of Health Care Organisations. The joint commission's second level of testing, beta testing, is a very extensive process involving hundreds of hospitals at a cost of more than $2000 each. Whilst the ACHS cannot currently test in this way, the ongoing analysis, review and modification of indicators following their implementation can be readily equated with the beta testing process.

In the context of validity, predictive ability and the strength of the empirical relationship between the clinical process and outcome should also be considered; some researchers refer to this as criterion validity. The question is: How valid are indicators in determining quality of care issues?

To highlight this point, reference is made to some guiding principles from the Maryland Hospital Association Quality Indicator Project in the United States (Kazandjian, Wood & Lawthers 1995). Principle 1 relates to the measurement of quality and states:

- indicators of performance do not measure quality, people do;
- indicators of performance may be measuring the quality of the data and not the goodness of care; and
- demonstrated usefulness is the best test of validity.

With this principle in mind, some researchers have suggested that the rate of unplanned readmissions, for example, is not a valid quality indicator because it fails to distinguish poor quality organisations from good quality organisations, and conclude therefore that the indicator is of little use. The Care Evaluation Program accepts that many factors, independent of the quality of care, may affect

an early readmission and maintains 'that the value of the indicator lies in its ability to stimulate an internal review of events' (Ansari, Collopy & Booth 1995). It is only through an internal clinical review that it can be determined if the readmission was unexpected and reflects a quality issue and not simply the progression of the disease process. It is the 'unexpected' unplanned readmissions which are reported in the Care Evaluation Program database.

This position is in keeping with that of the Maryland Association, whereby 'the Quality Indicator Project aims at the internal users of the information in their search for ways to improve performance and care. This approach has demonstrated its validity over time' (Kazandjian, Wood & Lawthers 1995).

Indicator responsiveness

The best test of indicator validity is its demonstrated usefulness and this is being determined through analysis of qualitative information. Australian health care organisations are responding well to indicator monitoring and their use continues to stimulate a variety of quality activities and to induce changes in clinical and administrative practice. Indeed, this is the focus of the new ACHS Evaluation and Quality Improvement Program – to demonstrate how information and data are being used within organisations to improve performance. This point is illustrated by the action taken as a result of clinical indicator monitoring. For all HWMIs addressed by organisations during 1995, 27 per cent of respondents have reported acting upon their data, as shown in Figure 1. This has mostly involved the initiation of further quality activities (30 per cent), whilst policy and procedural changes have collectively accounted for 35 per cent of actions taken. The other category (31 per cent) mostly relates to in-service education. Results

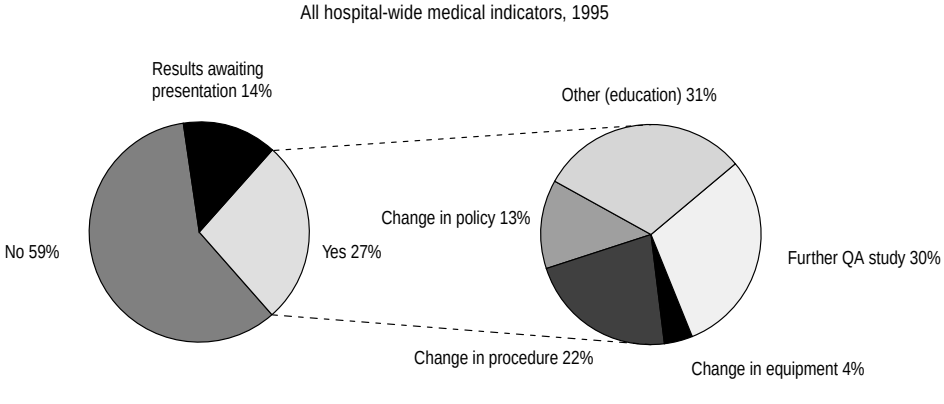

Figure 1: Action taken as a result of indicator monitoring

such as these are encouraging, given that in every instance where action is taken (and over 500 were reported for 1995) the potential exists to avoid an adverse event and to improve patient care.

For example, one organisation found that 17 per cent of day procedure patients 'failed to arrive'. In reviewing the data relating to this indicator, the patients were identified by the hospital as mostly being endoscopy patients. As a consequence, pre-admission procedures to confirm patient attendance were instigated and patient information brochures were established, with a reported reduction by the hospital in the 'failure to arrive' rate. In another example, only 20 per cent of patients were identified as receiving thrombolytic therapy within one hour of presenting to hospital with an acute myocardial infarction. Following internal review of this process, treatment protocols were subsequently modified to allow for the initiation of therapy within the emergency department, rather than waiting for patient admission to coronary care. In a third example, a hospital obtained a post-operative wound infection rate below threshold and queried whether, in fact, all relevant cases were being recruited. Following re-collection of the data, accuracy was confirmed, but new wound evaluation forms were subsequently developed for more thorough and ongoing use throughout the hospital.

These few examples are representative of the many positive outcomes that are increasingly being reported to the Care Evaluation Program by health care organisations and augur well in demonstrating the usefulness of indicators within internal quality programs. A separate paper on hospital responses for 1995 will be published shortly (Portelli, Williams & Collopy 1997).

Indicator reliability

The reliability and reproducibility of clinical indicator data (that is, the extent to which indicator occurrences will be accurately and completely identified from at risk populations) are of considerable importance to the integrity of the Care Evaluation Program database. Indicator results may vary as a result of data collection capabilities and hence the most fundamental question is: Are all participants reporting data in the same way?

There are a number of points for consideration including indicator and definitional precision, accuracy of data reporting and recording, and analysis of data elements to determine the extent of variation in results.

Reliable data collection will be enhanced by precision in defining the data elements and, as previously mentioned, the exclusion of data which are not consistent with given definitions. For example, the HWMI for post-operative

wound infection, relates to *all* surgical patients remaining in hospital on the fifth post-operative day (ACHS 1995). Precision will, however, be much greater for discipline-specific indicators, such as the surgical indicator relating to wound infection following total hip joint replacement, where just one category of patient is being reviewed (ACHS 1996).

There are many definition and rate construction methods to quantify nosocomial infections, evidenced by the degree of dispute over the classification of surgical incisions. However, by defining the method and excluding data obtained through the use of alternative methods, the integrity of the database can be maintained.

Indicators will also perform variably as a result of data collection capabilities, particularly if there is reliance on retrospective record review as the principal data source, combined with the influence of observer impression of the event in question; hence reiterating the need for clear definitions and consistent education. Data collection accuracy does appear to be influenced by the significance of the event in the eyes of the recorder. For example, in a study relating to the recording of events in the record by hospital medical officers, less than 25 per cent of instances of phlebitis were recorded, however, this rose to 80 per cent for instances of wound infection and 100 per cent for mortality (Collopy 1980). In support of this concept, in developing clinical indicators for monitoring with college working parties, a focus upon major or critical events is strongly advocated.

Another consideration is accuracy of coding and generally the standards of coding in Australia are high. In one study, coding accuracy by medical record administrators was demonstrated to be 95 per cent (Collopy 1980), whilst another re-coding study of hospital medical records in Victoria determined that 86.5 per cent of records were coded correctly (Department of Health and Community Services 1995). Nonetheless, some inaccuracies are likely, particularly if, as mentioned before, data are accumulated in a retrospective way. Failure to record events in the medical record and inaccuracy in coding may compound and influence the value or accuracy of administrative databases. This is an argument against the use of such databases for inter-hospital comparison of quality of care.

The size of the aggregate database is now quite substantial. As shown in Table 3, for three HWMIs after the first three years, the denominator figures (patients at risk for the event) approximate between 0.5 and 1.0 million patients. As the size of the database increases as anticipated, the variable performance of indicators due to data collection inaccuracies will have less influence upon the aggregated results.

Table 3: HWMI aggregate data, 1993–1995

Indicator	Numerator	Denominator	Rate (%)
Unplanned readmission	22 891	863 522	2.65
Return to operating room	3 378	577 765	0.58
Hospital-acquired bacteraemia	2 078	890 804	0.23

The extent of variation in reported data is assessed by the Care Evaluation Program through the application of various distributional statistics such as mean rates, confidence intervals, standard errors and co-efficient estimates.

In considering the mean rates of aggregated data for the three years, 1993 to 1995, there is remarkable consistency of results. This is particularly evident for those indicators regarded by some as being less than reliable and subject to observer bias, such as unplanned readmissions and unplanned returns to the operating room (Table 4). This consistency is achieved despite the fact that the data accumulated by the Care Evaluation Program for each year have been, to date, from differing hospitals all over Australia, in the three-year cycle by which the majority of hospitals have been surveyed.

Table 4: HWMI mean rates, 1993–1995

Indicator	1993	1994	1995
Pulmonary embolism	0.53	0.20	0.27
Unplanned readmission	3.1	2.3	2.7
Unplanned return to operating room	0.55	0.6	0.56
Clean wound infection	1.1	2.7	1.4
Contaminated wound infection	2.1	2.8	2.2
Hospital-acquired bacteraemia	0.09	0.28	0.26

In further considering the aggregated data for 1993 and 1994, there is a clear narrowing of the 95 per cent confidence intervals for the majority of the rate-based clinical indicators. This information is reported in some detail in the report, 'Measurement of care in Australian hospitals' (Ansari et al. 1995). Table 5 demonstrates this trend for the indicators relating to pulmonary embolism and unplanned readmission. These data provide some assurance of the precision of the indicators and the accuracy of the results that are being aggregated and presented to the health care industry.

Table 5: HWMI rates and confidence intervals

Indicator	Rate (%)	95% confidence interval
1993 data		
Pulmonary embolism	0.53	0.43–0.66
Unplanned readmission	3.10	2.7–3.7
1993–1994 data		
Pulmonary embolism	0.25	0.20–0.31
Unplanned readmission	2.60	2.30–2.90

Risk adjustment

The last area of consideration, referred to briefly, is that of risk adjustment. The crude comparison of indicator data across organisations does not account for differences. To 'level the playing field' so that rates can be compared, an appropriate adjustment method is required from a growing number of options, including casemix adjustment based on diagnosis related groups, illness severity indices and hospital characteristics. The merits of each of these have been explored in a previous paper (Booth, Ansari & Collopy 1995). In the absence of patient-specific data, the Care Evaluation Program is currently adjusting aggregate HWMI data according to hospital size, which has previously been shown to be a surrogate for casemix adjustment (Ansari, Collopy & McDonald 1996). The method employed is simple, easily understood, cost-effective and requires no additional data to be provided by health care organisations.

This method of adjustment is also being applied to stratify indicator thresholds as more data are accumulated. Thresholds are currently being established for three categories of hospital size: 1–99 beds, 100–199 beds and greater than 200 beds, and the values are based on two standard errors around the mean value. A clinical outcome that lies within the threshold range is considered acceptable practice. However, organisations should strive to achieve the best result in keeping with the intent of improving performance, which will generally be towards the lower end of the threshold range, depending on the intent of the indicator. This system has been well received so far by the industry and enables more meaningful information to be provided for comparative purposes. It is anticipated that similar adjustment methods will be applied to other sets of indicators as more data become available. However, the stratification variable used will depend upon the discipline under study, for example, for obstetrics

indicators, 'the number of mothers delivering', and with anaesthetic indicators, 'the number of procedures with an anaesthetist in attendance'. This may provide a more acceptable peer organisation than the number of hospital beds. So far, however, the results do not differ from those obtained through stratification by size.

Finally, many of the individual discipline (college-developed) indicators are either risk-adjusted in their format, for example, mortality for coronary artery bypass grafts with seven subsets, or by their nature, for example, anastomotic breakdown after surgery for colon cancer.

Conclusion

The Care Evaluation Program recognises that the National Aggregate Database has both strengths and weaknesses, some of which have been highlighted through this paper. Strengths include the size of the database, face and content validity of the measures, indicator responsiveness and reliability of the measures, particularly as internal monitoring tools. It is recognised that the database also has limitations, including the specificity of some indicators, data collection accuracy issues and limited risk adjustment models, given the absence of patient-specific data. However, as stated by O'Leary (1995), 'we should not expect to get performance measurement right immediately. This is going to be an evolutionary process and those who are to be measured must be engaged in the continuing development process'. The ACHS Care Evaluation Program looks forward to continuing to work with the Australian health care industry in further developing and improving its National Aggregate Database.

References

Ansari M, Brosi J, Booth J & Collopy B 1994, *Measurement of Care in Australian Hospitals – 1994 Hospital Wide Medical Indicator Results*, Australian Council on Healthcare Standards, Sydney.

Ansari M, Collopy B & Booth J 1995, 'Hospital characteristics associated with unplanned readmissions', *Australian Health Review,* vol 18, no 3.

Ansari M, Collopy B & McDonald I 1996, 'Establishing thresholds for adverse patient outcomes', *International Journal for Quality in Health Care*, vol 8, no 3, June.

Audet A & Denman Scott H 1993, 'The uniform clinical data set', *Annals of Internal Medicine*, vol 119, no 12, pp 1209–13.

ACHS (Australian Council on Healthcare Standards) 1995, *Clinical Indicators. A Users' Manual, Hospital Wide Medical Indicators*, Version 4.

ACHS (Australian Council on Healthcare Standards) 1996, *Clinical Indicators. A Users' Manual, Surgery Indicators*, Version 1.

Booth J, Ansari M & Collopy B 1995, 'Alternatives to measuring outcomes and quality through casemix', paper presented at National Casemix Conference, August.

Collopy B 1980, *Surgical Audit: A Pilot Study*, St Vincent's Hospital, Melbourne.

Department of Health and Community Services 1995, *Recoding Study of Hospital Medical Records,* May.

Kazandjian V, Wood P & Lawthers J 1995, 'Balancing science and practice in indicator development', *International Journal for Quality in Healthcare*, vol 7, no 1, pp 39–46.

Keogh G 1996, 'Surgical site infection – Validation of data collection', *RACS Annual Scientific Congress*, Abstract booklet, p 404.

O'Leary D 1995, 'Measurement and accountability: Taking careful aim', *Journal of Quality Improvement*, vol 21, no 7, July.

Portelli R, Williams J & Collopy B 1997, 'Using clinical indicators to change clinical practice', *Journal of Quality in Clinical Practice*.

Rose G & Barker D 1986, *Epidemiology for the Uninitiated,* British Medical Journal, 2nd edn, London.

Sibbritt D 1995, 'Validation of a 28 day interval between discharge and re-admission for emergency re-admission rates', *Journal of Quality in Clinical Practice,* vol 15.