Original article

doi:10.1111/j.1463-1318.2012.03143.x

Cancer auditing – how accurate are your data?

C. Platell*† and C. Penter*

*Colorectal Cancer Unit, St John of God Hospital, Subiaco, WA, Australia and †School of Surgery, University of Western Australia, Perth, WA, Australia

Received 12 March 2012; accepted 10 May 2012; Accepted Article online 25 June 2012

Abstract

Aim Human involvement in the collection and entering of information into a database leads to a degree of error. The aim of this study was to assess the concordance between two individuals blinded from each other who independently collected information on the same set of patients and entered it into a colorectal neoplasia database.

Method A colorectal research nurse and a surgeon independently maintained an electronic database on all new patients admitted with colorectal neoplasia under the surgeon over a 5-year period. Twenty-three key endpoints were selected from the database in order to determine the agreement between the two observers. The κ statistic (for nominal and ordinal data) and the concordance correlation coefficient (for interval data) were used to determine the level of agreement between the two data sets.

Results Both observers recorded 432 new referrals during this period. There was only complete concordance between the two databases with respect to the number of new patients and returns to theatre within 30 days. Nonetheless, there was almost perfect concordance between the two data sets for a majority of the endpoints. The most important areas of variance were in the length of stay ($\kappa = 0.78$), the American Society of Anesthesiology grade ($\kappa = 0.41$), emergency surgery ($\kappa = 0.36$), nodal staging ($\kappa = 0.54$) and time to recurrence ($\kappa = 0.77$).

Conclusion This study highlights a number of important areas of data inaccuracy in a prospective colorectal database. The inaccuracies were due to observer bias, issues of data interpretation, or just difficulty in collecting the information accurately.

Keywords Colorectal cancer, audit, kappa statistic

What is new in this paper?

This study highlights the errors in maintaining a cancer database and offers solutions to prevent them. It also demonstrates the impact these errors can have on subsequent analyses.

Introduction

There is currently a commitment to auditing surgical and cancer outcome [1,2]. This information can then be used for a surgeon's personal interest in his or her results or to guide national policy. How the information is collected varies between units, hospitals and healthcare systems. Considering its importance, it is interesting to note the limited amount of literature to determine how reliable these data are. Such information relies on individuals' interpreting and entering data into a database and any system that includes a human link will contain an element of error.

There are ways of constructing a database to try and reduce these errors, but they can only correct certain problems. Our group was interested to determine the variance in results of two individuals entering independently collected data relating to the same patients into the same type of database over the same period of time. Neither group could be considered a 'gold standard' as both the individuals recording the data were applying their own interpretations of the information that they were presented with. They would also make their own errors when entering the information.

The aim of this study was therefore to assess the concordance between the data entry of two individuals collecting information for a standardized database on the same set of patients with colorectal neoplasia over a 5-year period.

Correspondence to: Prof. Cameron Platell, St John of God Hospital, Subiaco, Colorectal Cancer Unit, 212/25 McCourt St, Subiaco 6008, Australia.
E-mail: cameron.platell@uwa.edu.au

© 2012 The Authors

Methods

Colorectal neoplasia was classified according to the American Joint Committee on Cancer guidelines [3]. A colorectal research nurse (CRN) commenced collection of data on patients with colorectal neoplasia who were managed by a group of four surgeons at St John of God Hospital Subiaco in July 2006. At the commencement of the study period, one of the surgeons proposed to prospectively study the accuracy of the data that were collected by maintaining an identical database for his own patients over a 5-year time period. The nurse was blinded to the fact that the surgeon was collecting the information. The database was written in Filemaker Pro 11.0 (FileMaker Inc., Santa Clara, California, USA) and had a designated dictionary to define relevant terms. CRN collected the information on a paper form and then transferred it to the computerized database twice a week.

The database included double entry of some important data-points, and field limitations to prevent obvious errors in data entry. The database had previously been in use for more than 10 years and had formed the basis for a number of publications [4]. The study was stopped in December 2011. A comparison was then made between the data collected by the CRN on the surgeon's patients and those maintained by the surgeon on the same group of patients. A comparison was made between 23 key endpoints in the two databases to determine the concordance between the two sets of observations. The endpoints were chosen as they were felt to reflect key pieces of information that would be important in subsequent analyses.

Statistical analysis

The two sets of data were exported into Stata12 (StataCorp LP, College Station, Texas, USA) for analysis. This analysis involved 23 endpoints that included data stored in either a nominal, ordinal or interval format. The κ statistic was used as a measure of agreement between the two observers for nominal and ordinal data and was scaled between 0 (implying that the amount of agreement is what would be expected by chance) and 1 (implying that there is perfect agreement between the observers). The relevance of the κ statistic is shown in Table 1 [5]. A concordance correlation coefficient was used to look at the association between observations for interval data (i.e. age, length of stay, height from anal verge for rectal cancer and time to recurrence) [6].

Results

Both databases recorded the same number and identity of patients over the 5-year study period (n = 432). There was only complete concordance (κ = 1.00) between the two sets of data for one of the 23 selected endpoints. This was the number of patients returning to theatre within 30 days of the initial surgery. There was almost perfect [5] concordance (κ = 0.81–1.00) for 14 of the 23 endpoints (Tables 2 and 3). For the remaining endpoints the degree of concordance varied from substantial [5] for six endpoints (κ = 0.61–0.8), moderate [5] for two endpoints (0.41–0.60) and fair [5] for one endpoint (0.21–4). As an example, the variation in the recording of binary information is detailed in Table 4.

Table 1 κ statistic: measure of agreement.

κ	Interpretation
0.0	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

The postoperative length of stay recorded by the two observers differed in a number of areas. CRN recorded the period as a mean of 8.4 ± 5.9 days (median 7 days, interquartile range 5–11, minimum 1 day and maximum 38 days) whereas the surgeon recorded the information as a mean of 7.4 ± 4.7 days (median 7 days, interquartile range 5–9, minimum 1 day and maximum 28 days). The principal area of error between the two observers was in patients staying for prolonged periods of time and who were subsequently transferred to either palliative or transitional care units. The surgeon only recorded the stay up to discharge from the surgical unit, whereas the CRN recorded the stay until the patient was discharged home.

There were also important differences in the recording of the American Society of Anesthesiologists (ASA) grade. This was used as a marker of the general fitness of the patient. CRN used the anaesthetic chart to record this, whereas the surgeon made his own independent assessment based upon the guidelines. CRN recorded ASA grade 1 33% (142), ASA grade 2 49% (211), ASA grade 3 16% (70) and ASA grade 4 2% (9), whereas the surgeon recorded ASA grade 1 30% (129), ASA grade 2 45% (193), ASA grade 3 22% (93) ASA grade 4 4% (17). Clearly, there was a strong bias for the surgeon to record the patients as having more ASA grades 3 and 4 compared with the CRN. In contrast, the surgeon recorded more episodes of recurrence of neoplasia (27 *vs* 21 patients) and more anastomotic leakages (8 *vs* 7) compared with the data collected by the CRN. The one leakage missed by the CRN was an asymptomatic

© 2012 The Authors

Table 2 The concordance (κ) between the various endpoints for the two databases (n = 432).

Data fields	Agreement (%)	κ	Standard error	P value
Sex	95.9	0.918	0.050	< 0.0001
ASA grade	61.81	0.410	0.033	< 0.0001
Emergency	88.7	0.361	0.048	< 0.0001
Cancer location	90.9	0.867	0.026	< 0.0001
Resection status	97.1	0.832	0.044	< 0.0001
Number of anastomoses	92.1	0.836	0.047	< 0.0001
Type of anastomosis	88.0	0.838	0.026	< 0.0001
G staging	83.5	0.659	0.044	< 0.0001
T staging	86.8	0.839	0.022	< 0.0001
N staging	74.4	0.544	0.035	< 0.0001
M staging	89.6	0.689	0.048	< 0.0001
TNM staging	83.3	0.790	0.024	< 0.0001
Nodes excised	95.4	0.952	0.013	< 0.0001
Involved lymph nodes	93.2	0.931	0.022	< 0.0001
Stoma creation	92.8	0.792	0.048	< 0.0001
Anastomotic leakage	99.8	0.932	0.048	< 0.0001
Return to theatre	100	1.000	0.048	< 0.0001
Chemotherapy	93.3	0.819	0.048	< 0.0001
Radiotherapy	96.8	0.878	0.048	< 0.0001

Table 3 Concordance correlation coefficient for integral data between the two databases.

Data fields	Concordance correlation coefficient	95% CI	P value
Age	0.999	0.999–0.999	< 0.001
Rectal cancer height	0.872	0.837–0.906	< 0.001
Postoperative stay	0.785	0.752–0.819	< 0.001
Time to recurrence	0.772	0.591–0.953	< 0.001

Table 4 A comparison of the difference between the two observers for various binary endpoints.

Endpoint	Colorectal research nurse (CRN) (%)	Surgeon (%)
Emergency surgery	38 (8.8)	47 (10.9)
Creation of end colostomy	93 (21.5)	98 (22.7)
Anastomotic leak	7 (1.6)	8 (1.8)
Return to theatre (30 days)	6 (1.4)	6 (1.4)
Chemotherapy commenced	111 (25.7)	102 (23.6)
Radiotherapy	70 (16.2)	66 (15.3)

radiological leak found on routine water-soluble contrast enema in a patient who had had a low anterior resection.

An analysis of the staging information demonstrated conflicting results. Some of this information was double entered to cross-check it. For example, there was almost perfect concordance between the two data sets with respect to the number of lymph nodes in the surgical specimen (κ = 0.95) and the number of those nodes that were positive (κ = 0.93). Nonetheless, the concordance was less accurate when comparing the actual nodal staging classification (N stage). The majority of these differences were in the recording of a patient as Nx (nodal status could not be accurately defined). This was recorded in 8% (25) of patients in the CRN database and in 13% (40) of patients in the surgeon's database. Similarly, the same error was noted in the recording of the metastatic staging (M stage). Here the surgeon's database was less likely to record a patient as Mx (seven (2%) patients) compared with the CRN database (21 (7%) patients).

Discussion

The study has shown good concordance between two data sets that were collected on the same group of patients with colorectal neoplasia over a 5-year period by two independent observers. Although many of the differences in the recorded information were minor, in some key areas they appeared to be significant. This variation could have an influence on endpoints such as the rate of local recurrence or disease-free survival. The most substantial disagreements were in the recording of information on the length of stay, ASA grade, the requirement for emergency surgery, the N staging classification and the development of recurrence. This

© 2012 The Authors

has relevance to how we validate and interpret information contained in clinical databases.

The reasons for the variations differed. For length of stay, the two observers used a different interpretation of the definition, despite what was recorded in the database dictionary. The surgeon recorded length of stay until the patient either went home or was transferred to another unit (i.e. palliative care or the transitional care unit for rehabilitation). The CRN recorded length of stay until discharge home. With the ASA grade, the CRN used the anaesthetic consultant's assessment whereas the surgeon made his own assessment. Clearly the surgeon demonstrated a bias to overestimate the ASA grade for the group of patients treated. With regard to the N staging, it was interesting to note there was high concordance in relationship to the number of nodes excised and the number of nodes involved with cancer. Nonetheless, when it came to the more interpretive step of entering the N stage, then the degree of concordance worsened.

In any database, some information is more easily obtained than other information [1,7]. In this study, the CRN under-reported patients presenting with recurrent disease, either local or systemic. This is information that comes from close follow-up of a cohort of patients. It relies on communication between the surgeon and the CRN to register this information. This is difficult information to capture and requires a great deal of vigilance to keep it accurate. It has been found that 'data fields that are not cognitively integrated with the other elements of the database' are more prone to error [1]. Recurrence and the date of last follow-up are examples of this point. This information is difficult to capture and maintain.

Several authors have noted variability in the accuracy of clinical databases. Lambert *et al.* [8] reviewed a hospital discharge database regarding cardiac comorbidities and found good agreement in the information. Any errors noted would have had no effect on 30-day and 1-year cardiac mortality predictions. In contrast, Wyse *et al.* [9] reviewed hospital endoscopy reports and found that physician claims underestimated the number of procedures performed by more than 15%. This raises the issue of what incentives there are to clinicians to over- or under-report information. Clearly in this study the surgeon demonstrated a bias to overestimate the sickness of the patients by recording high ASA scores.

Other groups have attempted to determine the reliability of information in their databases. Zhang *et al.* [10] compared a hospital cancer database with a personal clinician's database for 276 patients with breast cancer. They noted significant differences between the two databases with respect to the use of treatments such as radiation, chemotherapy and hormone therapy. However, the largest source of error was in the interpretation of the cancer T staging with a 28% error rate. Khan *et al.* [11] reviewed a large general practitioner database and noted major inaccuracies in the recording of acute conditions. This is similar to the problems noted in this study in the recording of emergency surgery information. Finally, Holt *et al.* [12] compared English hospital episode statistics with a gold standard data set for a sample of nine hospitals performing elective abdominal aortic aneurysm repairs over 2 years. The main outcome event was operative mortality and the authors found there was no difference between the two data sets. This is not at all surprising as mortality is a fairly easy endpoint to detect and such results would be expected. Nonetheless, when they examined comorbidity they noted significant differences between the databases.

It is difficult to maintain high quality data if several groups are involved or if one is trying to integrate data from multiple sources [7]. This is especially so if it is considered that many different members of the junior staff are involved in the data collection and entering process and that their motivation and attention to detail may vary. Similarly, they can only enter information they know, so if a surgeon were to withhold information then it would impact on this process.

The surgeon or oncologist probably remains in the best position to record and determine certain endpoints like local or systemic recurrence. Goldberg *et al.* [1] detected high error rates between several clinical oncology research databases. For example, an analysis of identical sets of tumour recurrence data on 133 patients found recurrence-free survival at 5 years was 69% for physician A and 61% for physician B. Goldberg *et al.* identified that many of these errors were due to mistakes in entry or to misinterpretation of information from the original documents. They identified that using alarms based on data constraints (e.g. age between 5 and 105 years) only seemed to prevent a small fraction of errors. Interestingly, many of the observed errors were non-random and were organized in clusters. It was almost as if the data entry officers were having a period of poor cognition, and if they made one error they seemed to make others. In this evaluation, Goldberg *et al.* identified three types of database error, i.e. that the original paper document was wrong and then copied to the database, or there was an error in interpretation of the data, or that there was simply an error in entering the data.

The database used in this study had built-in checks to avoid impossible or internally inconsistent data. Yet the use of such measures can only prevent certain types of error. The database also used double entry as a strategy to capture high value information such as staging. Yet, as

© 2012 The Authors

others have observed, this strategy is not error proof [1,7,13]. Therefore, in response to the differences seen between the two observers, a number of changes have been made in how the data are entered, checked and interpreted. The definitions for length of stay and recurrence were reworded so that there were no difficulties with interpretation. The recurrence section of the database was also reorganized so that it was clear (i) if a patient who was previously disease free developed recurrence, (ii) the date of recurrence, (iii) if the recurrence was local or systemic, (iv) if the recurrence was managed and (v) if the patient was currently disease free following retreatment. We have encouraged surgeons to discuss 'all' patients with recurrent disease at the weekly multidisciplinary team meeting. It was also determined that the ASA grade should be determined by the anaesthetist and recorded accordingly. In addition, we reviewed the classification of the N status and in particular the use of the Nx terminology to try and reduce future errors. This is information that needs to be double checked before entry. Finally, we will continue to collect this information independently of each other so as to cross-check every 6 months as part of an ongoing concordance study.

In summary, this study has highlighted a number of important areas of data error in a prospective colorectal cancer database. In particular, the problems of collecting recurrence data are most likely to be relevant to all surgeons who collect and maintain such databases. There are a number of recognized strategies that may improve data accuracy. These include the double entry of high value data, clinical alerts to restrict the numbers of patients in certain classifications (e.g. ASA 4) [13] and better training of staff to interpret information [1,7]. In high value fields like staging information for cancers, the database can include constraints to force the user to double check the entry, or to use a second person to double check all scoring and assessment tools. Regular reviews with team members are also recommended to ensure that problems with data collection and processing are identified early [7]. Yet a degree of interpretation is often required for the 'square peg in a round hole' problem where the patient's clinical details just do not seem to fit the database. The direct entry of data using electronic devices may also have an advantage over paper collection and then entry by removing a step in the process that has potential for error [14]. However, Azar *et al.* [15] identified that most of the inaccuracies in clinical databases are due to human error rather than system design.

In conclusion, this study has been a valuable exercise in trying to improve the accuracy of a clinical oncology database maintained at a single institution. It has highlighted a number of data fields that are open to error and interpretation or which are simply difficult to capture. This is particularly relevant to the staging information and the interpretation of the various T, N and M classifications. Human error is something we have to accept and recognize when managing any clinical database. Vigilance and double-checking and rechecking of information is essential.

References

1 Goldberg SI, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. *AMIA Annu Symp Proc* 2008; **2008:** 242–6.

2 Platell C. The secrets to a successful database. *ANZ J Surg* 2008; **78:** 729–30.

3 American Joint Committee on Cancer (2002 6th edition) *AJCC Cancer Staging Manual*. Springer, New York, pp. 113–24.

4 Platell C, Ng S, O'Bichere A, Tebbutt N. Changing management and survival in patients with stage IV colorectal cancer. *Dis Colon Rectum* 2011; **54:** 214–9.

5 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33:** 159–74.

6 Lin L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45:** 255–68.

7 Schmitt CP, Burchinal M. Data management practices for collaborative research. *Front Psychiatry* 2011; **2:** 47.

8 Lambert L, Blais C, Hamel D *et al.* Evaluation of care and surveillance of cardiovascular disease: can we trust medico-administrative hospital data? *Can J Cardiol* 2012; **28:** 162–8.

9 Wyse JM, Joseph L, Barkun AN, Sewitch MJ. Accuracy of administrative claims data for polypectomy. *CMAJ* 2011; **183:** E743–7.

10 Zhang M, Higashi T, Nishimoto H, Kinoshita T, Sobue T. Concordance of hospital-based cancer registry data with a clinicians' database for breast cancer. *J Eval Clin Pract* 2012; **18:** 459–64.

11 Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010; **60:** e128–36.

12 Holt PJE, Poloniecki JD, Thompson MM. Multicentre study of the quality of a large administrative dataset and implications for comparing death rates. *Br J Surg* 2012; **99:** 58–65.

13 Haerian K, McKeeby J, Dipatrizio G, Cimino JJ. Use of clinical alerting to improve the collection of clinical research data. *AMIA Annu Symp Proc* 2009; **2009:** 218–22.

14 Walther B, Hossin S, Townend J, Abernethy N, Parker D, Jeffries D. Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. *PLoS ONE* 2011; **6:** e25348.

15 Azar KM, Moreno MR, Wong EC, Shin JJ, Soto C, Palaniappan LP. Accuracy of data entry of patient race/ethnicity/ancestry and preferred spoken language in an ambulatory care setting. *Health Serv Res* 2012; **47:** 228–40.

© 2012 The Authors