Analysis of Data Errors in Clinical Research Databases

Saveli I. Goldberg, PhD[a], Andrzej Niemierko, PhD[a,d], Alexander Turchin, MD, MS[b,c,d]
[a]*Massachusetts General Hospital, Boston, MA*
[b]*Clinical Informatics Research and Development, Partners HealthCare, Boston, MA*
[c]*Brigham and Women's Hospital, Boston, MA*
[d]*Harvard Medical School, Boston, MA*

Abstract

Errors in clinical research databases are common but relatively little is known about their characteristics and optimal detection and prevention strategies. We have analyzed data from several clinical research databases at a single academic medical center to assess frequency, distribution and features of data entry errors.

Error rates detected by the double-entry method ranged from 2.3 to 26.9%. Errors were due to both mistakes in data entry and to misinterpretation of the information in the original documents. Error detection based on data constraint failure significantly underestimated total error rates and constraint-based alarms integrated into the database appear to prevent only a small fraction of errors. Many errors were non-random, organized in special and cognitive clusters, and some could potentially affect the interpretation of the study results. Further investigation is needed into the methods for detection and prevention of data errors in research.

Introduction

Errors are common in patient care[1, 2] and can lead to adverse events[3]. Importance of prevention of errors in clinical care is well recognized[4] and there is a large body of research investigating prevention strategies.

However, errors of direct care are not the only ones that can harm patients. Errors in clinical research, if large enough to affect the investigators' conclusions, can have much greater impact on clinical outcomes by swaying the standard of care of thousands of patients[5]. In fact, a number of reports have shown that errors are common in clinical research databases[6-9]. Nevertheless, relatively little is known about the types of errors in research databases, their characteristics and possible effects on research conclusions. We therefore undertook this project to examine prevalence and features of apparent errors in several clinical research databases.

Materials and Methods

Dataset

We analyzed the data from several research databases that contained information about treatment and outcomes of oncologic patients who underwent radiation treatment at a single academic medical center. The databases used MS Access client and PostgreSQL database server. Standard MS Access forms graphical user interface was used for data entry. All data in these databases were entered manually by trained technicians, usually being copied from electronic or paper medical records. Constraints by parameter-specific ranges and dynamic constraints based on values in other fields were used to minimize data entry errors. Individuals who entered specific records were not tracked. A typical record contained the patient's demographic information, date of diagnosis of their condition (defined as the date of biopsy), dates of initial and final outpatient radiation treatment visit, date of last follow-up visit (after the radiation treatment course had been completed), and current follow-up status (remission, relapsed, deceased from the treated cancer, deceased from other causes). We have employed two strategies for identifying erroneous entries: highly improbable / internally inconsistent data and data discrepancies between duplicate data entries in different databases (externally inconsistent data).

Impossible / Internally Inconsistent Data

To evaluate data in research databases for impossible entries and internal inconsistencies we analyzed two databases (subsequently referred to as "B" and "S") that contained data on treatment and outcomes of oncologic patients. Both databases contained similar data fields. However, while database B primarily contained information on patients who were diagnosed at the same hospital, database S contained a substantial fraction of patients who were diagnosed elsewhere and were subsequently referred for treatment.

In each of these databases we evaluated data for the following impossible conditions:

1. Date of diagnosis falls on a Sunday (date of diagnosis was defined as the date of the biopsy which are not normally conducted on weekends)
2. Date of the first radiation treatment falls on a Sunday (radiation treatments are usually only administered Monday through Friday)
3. Date of the last radiation treatment falls on a Sunday
4. Date of the last follow-up visit falls on a Sunday

We also analyzed the number of data entries that triggered data integrity alarms incorporated into the

databases. The alarms were triggered by the following impossible conditions:

1. Date of Diagnosis (database B only): triggered by date of diagnosis > date of the pathology report, date of diagnosis > date of initiation of chemotherapy, date of diagnosis > date of relapse, date of diagnosis > date of the last follow-up appointment.
2. Date of the first radiation treatment (both databases): triggered if < date of diagnosis, > date of last follow-up, > date of last treatment, > 3 months before the date of the last treatment (database B only: courses of radiation treatment for patients included in that database cannot be longer than 3 months)
3. Date of the last follow-up visit: triggered if < date of entry.

For both of these databases we also assessed internal consistency of the data on the example of concordance of the fields containing information about vital status and relapse status. These fields were considered internally inconsistent if vital status was recorded as "deceased from the cancer" but no relapse was documented for patients who were known to have gone into remission after conclusion of their initial course of treatment.

Externally Inconsistent Data

To analyze the data in research databases for external inconsistencies we analyzed 1,006 patient records that were incidentally entered in two different databases (subsequently referred to as "P1" and "P2") at the same time. We analyzed the discrepancies between the records of the same patients in the two databases in the following fields: medical record number (MRN), date of birth (DOB), first and last name, number of treatment sessions, and the dates of the first and last treatment session. All of the demographic information fields were entered on one screen in both databases, and all of the information related to treatment was entered on another screen.

In addition to analyzing discrepancies between individual fields in the two databases, we also analyzed concordance between discrepancies in the fields entered on the same screen and the fields entered on different screens.

To demonstrate a potential effect of errors in research data we also analyzed for mutual consistency two datasets on local tumor recurrence in 133 patients that were independently entered by two physicians. We assessed the differences in time to recurrence derived from each of these two datasets (which should have been completely identical).

Statistical Analysis

Binominal distribution was used to calculate exact 95% confidence limits for error frequencies. Fisher's Exact Test was used for analysis of 2x2 tables. Survival curves were compared using a log-rank test. All analyses were performed in SAS software program (Version 8.1;SAS, Cary, NC). All statistical tests were 2-sided.

IRB

The study protocol was reviewed and approved by Partners Human Research Committee.

Results

Databases B and S contained records of 5,859 and 2,520 patients, respectively (Table 1). Fraction of events on Sundays ranged from 0.48% for last treatment visit in database B to 2.34% for the date of diagnosis in database S. Rates of errors were similar in the same fields in both databases with the exception of the fraction of dates of diagnosis that fell on a Sunday that was substantially higher in database S compared to the database B (2.34 vs. 0.99%; $p < 0.0001$). The fraction of Sundays was significantly higher for the dates of the last follow-up visit (> 2% for both databases) than for the dates of first and last treatment ($p < 0.0001$ for database B and $p < 0.0001$ for Radiation Start Date and $p = 0.0031$ for Radiation End Date for database S).

The fraction of data entries that initially triggered data integrity alarms ranged from 0.2% of the dates of the last follow-up visit in database B to 1.9% of the dates of the first radiation appointment in database S (Table 2). There were no significant differences in alarm rates between the two databases.

Analysis of the vital status data in these databases showed 1,161 patients in the database B who achieved remission but had vital status "Deceased from the Treated Cancer". Of these, 98 (8.4%) did not have any information about relapse recorded. Similarly, 62 (10.6%) out of 584 patients in database S who had achieved remission and had vital status "Deceased from the Treated Cancer" did not have any information about disease relapse recorded.

Analysis of the duplicate data on 1,006 patients entered into two databases showed that rates of discrepancies between the two databases ranged between 2.3 and 5.2% for demographic data and between 10.0 and 26.9% for treatment data (Table 3). The rate of impossible values was similar to the databases B and S: 0.8% of dates of the first treatment in the database P2 fell on a Sunday.

Frequency of discrepancies in any field was higher if there was a discrepancy in the same patient record on another field on the same screen. Out of the 21 patients who had a discrepancy in MRN, 5 (23.8%) also had a discrepancy in the DOB. On the other hand, out of 67 patients who had a discrepancy in the number of treatment sessions, only 4 (5.97%) also had a discrepancy in DOB (p=0.03).

Table 1
Impossible / Internally Inconsistent Data Entry in Two Research Databases

	Database B		Database S		
Field	**Records**	**Sunday, N (%; 95% CI)**	**Records**	**Sundays, N (%, 95% CI)**	**P-value**
Diagnosis date	5859	58 (0.99; 0.75-1.28)	2050	48 (2.34; 1.73-3.09)	<0.0001
Radiation Start	3145	19 (0.60; 0.36-0.94)	1841	11 (0.60; 0.24-0.96)	1.0
Radiation End	3114	15 (0.48; 0.27-0.79)	1809	17 (0.93; 0.55-1.50)	0.065
Last Follow-up Visit	3696	77 (2.08; 1.65-2.60)	3006	61 (2.03; 1.56-2.60)	0.931

Table 2
Internal Data Integrity Alarms in Two Research Databases

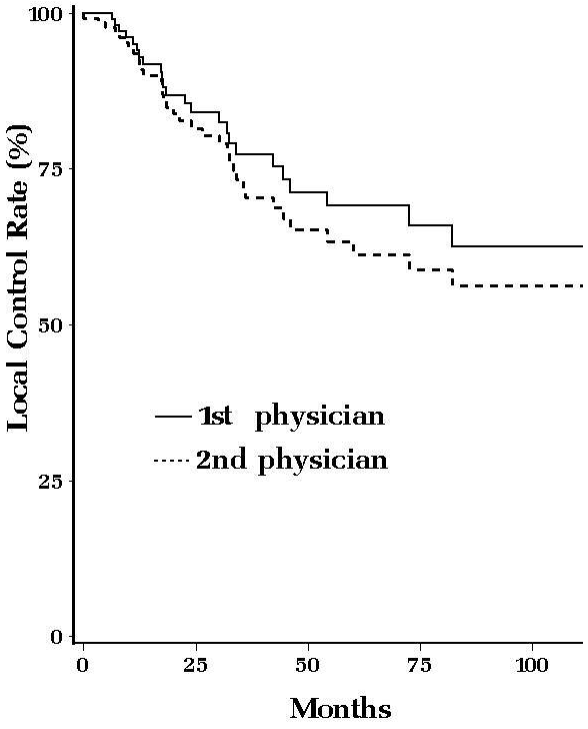
	Database B		Database S		
Field	**Records**	**Alarms, N (%; 95% CI)**	**Records**	**Alarms, N (%, 95% CI)**	**P-value**
Diagnosis date	5859	63 (1.08; 0.83-1.37)	2050	N/A	
Radiation Start	3145	40 (1.28; 0.91-1.73)	1841	35 (1.9; 1.33-2.63)	0.091
Last Follow-up Visit	3696	7 (0.2; 0.08-0.39)	3006	9 (0.3; 0.14-0.57)	0.452

Analysis of the discrepancies between two identical data sets on local tumor recurrence on 133 patients showed a considerable trend towards a difference in time to recurrence (Figure 1). At the end of the five-year follow-up period, the data entered by one physician showed 69% (95% CI ± 5.8%) of recurrence-free survival while the data entered by the other physician showed 61% (95% CI ± 5.7%; p = 0.38). The errors that resulted in this difference between the two datasets included 18 (13.5%) records with missing or different diagnosis dates, 1 (3.7%) records with different local failure date and 9 (25.0%) records with discrepancies in failure type (local vs. distant) or missing relapse dates.

Table 3
Database "P1", "P2" for the same patients

Field	N	Discrepancies, N (%, 95% CI)
MRN	1006	23 (2.29; 1.45-3.41)
Last name	983	45 (4.58; 3.71-6.66)
First name	983	32 (3.26; 2.24-4.56)
DOB	868	45 (5.18; 3.81-6.88)
Treatment Sessions	668	67 (10.0; 7.9-12.6)
First Treatment Date	983	264 (26.9; 24.1-29.8)
Last Treatment Date	944	182 (19.3; 16.8-21.9)

Figure 1
Local Control Rate (2 different physicians)



Discussion

In this large analysis of several clinical research databases we found that errors in the data were common, including both incorrect and missing information. The rates of discrepancies between data fields entered in duplicate in two different databases were as high as 27%, corresponding to a 13.5% error rate in each of the databases.

Data errors in research databases can have several etiologies: a) errors originating in the initial documents that were subsequently copied into the database b) errors of interpretation of the data in the initial documents and c) errors of data entry into the database[10, 11]. Our analyses could only detect the

latter two categories; therefore the error rates we found are likely underestimates.

The error rates in the same data categories were similar across different databases. Of the four date fields in the databases B and S, only date of the initial diagnosis (biopsy) had significantly different frequencies of impossible entries (Sundays). That difference might have been due to the fact that more patients in the database S (which had a higher rate of Sunday diagnosis dates) were referred from outside healthcare facilities leading to a higher rate of errors in interpretation of the original data.

Integrity checks, which identify impossible or internally inconsistent data entries, are a common way to assess data for errors[12]. However, they only evaluate the data for a limited number of conditions and therefore are expected to underestimate the error rate. For example, identification of all appointment dates that fall on a Sunday would be expected to provide 1/7 of the actual error rate in the appointment dates. Our results corroborate this prediction: in the database P2 the rate of Sunday appointment dates for the first treatment visit was significantly smaller (0.8%) than the overall rate of discrepancies between P1 and P2 on this field (26.9%, suggesting 13.5% error rate in each of the databases). In fact, the difference was more than twice the 7-fold prediction that was based on the assumption that all date errors happen with the same frequency. It is therefore possible that the errors are not completely random but are, for example, more likely to fall on a date next to the date of the actual visit. Since there are no appointments on Saturdays either, this would lead to the frequency of Sunday appointments of about half of the frequency of erroneous entries that fall on Tuesday through Thursday.

Similarly to the retrospective integrity checks, alarms based on impossible / inconsistent data values would also be expected to prevent only a minority of erroneous data entries. Consistent with this expectation, the rate of recorded alarms in our data was similar to that of the retrospective data constraint failures and substantially lower than the rate of discrepancies between two fully identical databases.

Based on our results, there are some conditions that could lead to higher data error rates. Data fields that are not cognitively integrated with the other elements of the database appear to be more prone to errors. For example, the date of the last follow-up visit in databases B and S is informationally isolated from the other data, while the dates of the first and last treatment visit are tied together with other data into a cognitive treatment model. Consistent with this hypothesis, the error rates in the date of the last follow-up visit were more than two-fold higher than the error rates in either first or last treatment visit dates.

The data errors appear to be clustered in accordance with the spatial arrangement of the database fields in the data entry forms. Our evaluation showed that presence of one data error on the demographic information screen increased the probability of another data error in another field on the same screen several fold while no association was found between errors in the fields on different screens. One possible explanation could be that a single distracting event may be responsible for both errors on the same data entry screen but does not carry over to another screen.

The comparison between the survival results derived from the same dataset entered independently by two different physicians illustrates the risks of misinterpretation of the results of research brought upon by the errors in the data. While the difference in outcomes did not reach statistical significance in our example, the number of patients involved was small and a larger sample size could lead to this visually apparent divergence reaching the significance threshold. Though some authors have argued that mistakes in the data are unlikely to affect data interpretation[10], our results show that many of the errors are non-random and could therefore skew the final outcome.

An obvious method for reduction of data entry errors would be minimization of manual data entry by using direct data transfers from electronic medical records into research databases. However, this is not always possible. Manual data entry remains common in prospective studies where data is generated for the study itself rather than for clinical care (and is therefore not recorded in the electronic medical record). Another common scenario that mandates manual data entry involves retrospective data collection that requires cognitive synthesis of the data available in the medical record and / or abstraction of information from narrative medical documents that is not available as an exportable structured data field.

Our study points to several potential strategies for mitigation of data entry errors. Given that current integrity checks cover only a small fraction of potential errors, the number of constraints on data fields could be increased. While the number of absolute constraints that could be imposed is limited, partial constraints (which force the user to double-check the entry) could be employed. More extensive cognitive integration of data fields would likely also lead to a reduction in error rates, in effect imposing dynamic constraints that vary based on the context of the other fields in the record. In high value data entry double-entry or other techniques, such as read-aloud data entry could be employed[13], though expense

associated with their implementation may be substantial. Adequate training of the study stuff could ameliorate the interpretation of information during the data entry process[14].

Our study has a number of strengths. It involved multiple research databases and diverse strategies for error detection including both data constraints and double data entry, currently the gold standard. This comprehensive approach allowed us to compare different strategies for error detection and make projections about the rate of undetected errors. We were also able to demonstrate how data errors could potentially affect research results – an issue of vital importance, potentially affecting treatment and outcomes of thousands of patients.

Our study had several limitations. While to the best of our knowledge the data only included outpatient visits that do not take place on weekends, it is possible that in fact some of the diagnosis dates and initial treatment dates reflected emergency care delivered in the inpatient setting. However, our data were internally consistent with similar rates of Sunday dates between initial and last treatment visit (which is always outpatient) and expected ratios between frequency of errors detected by single-constraint and double-entry methods. Therefore, even if some of the care recorded in the database took place in the hospital, the frequency of this event was low enough not to affect our results. This preliminary study did not include a detailed analysis of the root causes of the data entry errors. The data analyzed was generated in the Department of Radiation Oncology in a single academic medical center and may not be applicable to other settings.

Conclusion

In this large study of data errors in several clinical research databases that the errors in research data are common, frequently non-random and only a minority of them can be stopped by the typically applied data constraint methods. These errors can potentially affect interpretation of research results. Further investigation is needed into the optimal approaches for detection and prevention of data errors in research databases.

References

1. Lesar TS, Briceland L, Stein DS. Factors related to errors in medication prescribing. *Jama.* Jan 22-29 1997;277(4):312-317.

2. Leape LL, Bates DW, Cullen DJ, et al. Systems analysis of adverse drug events. ADE Prevention Study Group. *Jama.* Jul 5 1995;274(1):35-43.

3. Bates DW, Boyle DL, Vander Vliet MB, Schneider J, Leape L. Relationship between medication errors and adverse drug events. *J Gen Intern Med.* Apr 1995;10(4):199-205.

4. Weingart SN, Wilson RM, Gibberd RW, Harrison B. Epidemiology of medical error. *Bmj.* Mar 18 2000;320(7237):774-777.

5. Goldhill DR, Sumner A. APACHE II, data accuracy and outcome prediction. *Anaesthesia.* Oct 1998;53(10):937-943.

6. Shelby-James TM, Abernethy AP, McAlindon A, Currow DC. Handheld computers for data entry: high tech has its problems too. *Trials.* 2007;8:5.

7. Beretta L, Aldrovandi V, Grandi E, Citerio G, Stocchetti N. Improving the quality of data entry in a low-budget head injury database. *Acta Neurochir (Wien).* 2007;149(9):903-909.

8. Seddon DJ, Williams EM. Data quality in population-based cancer registration: an assessment of the Merseyside and Cheshire Cancer Registry. *Br J Cancer.* 1997;76(5):667-674.

9. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc.* Nov-Dec 2002;9(6):600-611.

10. Day S, Fayers P, Harvey D. Double data entry: what value, what price? *Control Clin Trials.* Feb 1998;19(1):15-24.

11. van der Putten E, van der Velden JW, Siers A, Hamersma EA. A pilot study on the quality of data management in a cancer clinical trial. *Control Clin Trials.* Jun 1987;8(2):96-100.

12. Rahm E, Do HH. Data Cleaning: Problems and Current Approaches. *Bulletin of the Technical Committee on Data Engineering.* 2000;23(4):3-13.

13. Kawado M, Hinotsu S, Matsuyama Y, Yamaguchi T, Hashimoto S, Ohashi Y. A comparison of error detection rates between the reading aloud method and the double data entry method. *Control Clin Trials.* Oct 2003;24(5):560-569.

14. Lorenzoni L, Da Cas R, Aparo UL. The quality of abstracting medical information from the medical record: the impact of training programmes. *Int J Qual Health Care.* Jun 1999;11(3):209-213.