MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Data Entry Errors in Clinical Research Databases

Saveli Goldberg, Marek Ancukiewicz, Andrzej Niemierko

Massachusetts General Hospital, Boston, MA, USA

Abstract and objective

Human-caused errors occur at all stages of research process. We consider here data entry errors in clinical databases, understood as unintentional mistakes in interpretation or transcription of medical records at the time of data entry, and mistakes resulting from interaction with a computer program. Our experience shows that errors of this type occur commonly The existence, and even the consideration of a possibility of errors of data entry is seldom acknowledged in research practice, despite that they add substantially to the uncertainty about study results and, sometimes, may affect the conclusion of the study. We consider some systematic strategies of database errors reduction.

Keywords:

databases, factual

Methods

To investigate the prevalence of database errors and their characteristics we 1) tested 20 volunteers entering the same data sample, representative of Radiation Oncology databases and 2) used data from a review and consistency checks in a real database of 5500 patients. The volunteer test was performed using proprietary computer software for data entry, which enabled us to capture mistakes, self-corrections, and temporal characteristics of keystrokes. The computer program did not provide the user with a feedback about inconsistent entries, or exceeding allowed ranges for numeric data. However, text data were not allowed in numeric fields and conversely. Test data included 139 records and eight fields: three fields with character data in English text and five fields with numeric data. Between 34 and 100 records were entered per volunteer. A database containing information about 5500 patients was used for estimation of data entry errors. For real database, we considered as an evidence of data entry mistakes the treatment initiation dates and the follow-up date falling on Sunday. To illustrate the significance of database errors we derived two Kaplan-Meier estimates of local control obtained from data interpreted and entered by two physicians from the same medical records of 120 patients.

Results

In data entry experiments, an uncorrected mistake in was made in 0.2%-6.5% of fields entered, depending on the person tested. The average error rate was 2.7%, with S.D. of 2.0%. Type and percent of errors are relatively stable parameters for each person. The actual average distance between errors was significantly less then if we had a uniform distribution of errors. The probability of error in a field was found to depend on nature of the field and relationship of that field with other data. For example in database from 5500 patients Sunday as the day of radiation therapy start was in 0.9% cases, but Sunday as the day of last follow-up visit was in 2.6% cases. For two versions of the data prepared by two physicians, the actuarial local control gave the 69% +/- 6% vs. 61% +/- 6% in 5-year rates. Missing data about local failure is the main source of the survival curves distinction**.**

Conclusions

Database mistakes are a ubiquitous part of the normal research process. As we have demonstrated on Kaplan-Meier estimates of local control, data entry errors may have a substantial impact on study findings and the significance of the results. Different approaches for reduction of data entry errors include proper management of data entry process (including double data entry) and algorithms for detecting inconsistent and out of range entries during time of data entry. Further research is needed to develop algorithms for detecting data entry errors based on temporal characteristics of typing, self-correction of the record, history of the user previous errors, and the style of user typing. A caution should be always exercised during statistical analysis of data and in interpreting study findings. It is possible to propose some methods of database errors reduction, which take into account the statistical properties of the mistakes.