Case Report ■

Reducing Errors from the Electronic Transcription of Data Collected on Paper Forms: A Research Data Case Study

MONIKA M. WAHI, MPH, DAVID V. PARKS, BSEE, MBA, ROBERT C. SKEATE, MD, STEVEN B. GOLDIN, MD, PHD

Abstract We conducted a reliability study comparing single data entry (SE) into a Microsoft Excel spreadsheet to entry using the existing forms (EF) feature of the Teleforms software system, in which optical character recognition is used to capture data off of paper forms designed in non-Teleforms software programs. We compared the transcription of data from multiple paper forms from over 100 research participants representing almost 20,000 data entry fields. Error rates for SE were significantly lower than those for EF, so we chose SE for data entry in our study. Data transcription strategies from paper to electronic format should be chosen based on evidence from formal evaluations, and their design should be contemplated during the paper forms development stage.

■ **J Am Med Inform Assoc.** 2008;15:386–389. DOI 10.1197/jamia.M2381.

Introduction

Transcription of data from paper forms into an electronic database can be a nontrivial source of error.[1–3] Despite this fact, and often for valid reasons, data collection is often initiated on paper forms without consideration given to how the data will be transferred from paper to electronic format. Paper data collection forms are often developed in word processing programs such as Microsoft (MS) Word.[4] Researchers and health care providers are generally comfortable with these programs, whereas the realm of database design and programming using MS Access[5] or more enterprise structured query language (SQL) products such as Oracle SQL[6] and MS SQL Server[7] often lies outside of their training and expertise. Paper forms can be developed for a grant or Protection of Human Subjects application. Thus, designing an entire data entry system at this stage would seem to represent a poor use of resources.

Because data collection for health care research projects often begins with paper forms and these forms tend to pile up quickly, it is tempting to look at simple and time-honored approaches to electronic transcription such as duplicate data entry (DE) to protect data quality. In DE, two different individuals enter the same data into two different datasets, the datasets are compared electronically, discrepancies are flagged as errors, and these errors are manually resolved.

Studies of DE and single data entry (SE) have provided estimates of error rates. The SE error rates can be quite variable, and have been reported to be as low as 10.8[8] and as high as 124 per 10,000 fields.[3] In one study where two SE datasets were created from the same data, 6.5% of the entered fields did not match in the two datasets. This translates to an error rate of 650 per 10,000 fields.[9] In a study where two professional data managers conducted SE and consistency checks, error rates were lower, at 13 and 15 errors per 10,000 fields, the lower rates being attributed to the addition of consistency checks.[10] Although there are few studies on DE, one study compared DE and SE and found that DE reduced the error rate from 22 to 19 per 10,000 fields.[11]

DE includes some important positive features, including the ability to quickly implement the data conversion process, to use junior staff with no database design expertise, to use standard spreadsheet software such as MS Excel,[12] and to run simple queries (such as Proc Compare[13]) to identify discrepancies (errors) between the spreadsheet datasets. The main drawback to DE is the amount of labor required. DE is best accomplished by dedicating two separate individuals to the data entry task, but this may be beyond the organization's resources. Further, a technically trained data manager needs to be involved to compare the spreadsheets and manage the data quality, and this often proves to be more time consuming than anticipated at the outset.

The purpose of this report is to describe our research group's approach when faced with the prospect of accurately trans-

Affiliations of the authors: Department of Epidemiology and Biostatistics, University of South Florida College of Public Health (MMW), Tampa, FL; Department of Facilities and Academic Support for Technology, Johnnie B. Byrd, Sr., Alzheimer's Center and Research Institute (MMW, DVP), Tampa, FL; North Central Blood Services, American Red Cross (RCS), St. Paul, MN; Department of Surgery, University of South Florida College of Medicine (SBG), Tampa, FL.

Supported by the Johnnie B. Byrd, Sr., Alzheimer's Center and Research Institute.

The authors thank Jonathan McKeithan for conducting the single data entry described in this article and Keir Bradshaw for creating the figures.

Correspondence: Monika M. Wahi, MPH, Department of Facilities and Academic Support for Technology, Byrd Alzheimer's Institute, 4001 East Fletcher Avenue, Tampa, FL 33613; e-mail: <mwahi@byrdinstitute.org>.

Received for review: 01/17/07; accepted for publication: 01/02/08.

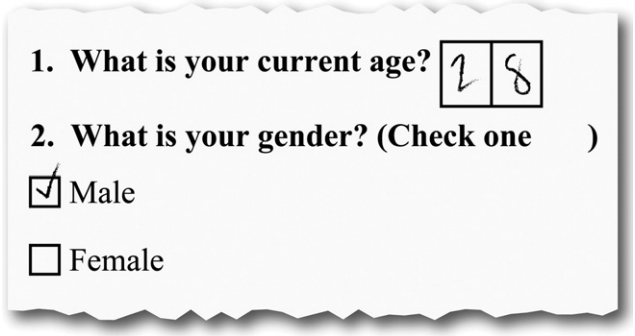

Figure 1. Example of Teleforms traditional forms.

ferring data from manually completed paper forms into electronic format with no precontemplated data entry system in place and minimal resources available.

Case Description

Paper questionnaires from the Surgical Clerkship Study (SCS),[14,15] an ongoing study of medical students at the University of South Florida College of Medicine, were designed in MS Word[4] and had been completed by over 100 students before a data entry system was selected. SE or DE systems were the first to be considered. Because only a small amount of time from a junior staff member and from a data manager was available to maintain a data system, SE was initially favored. However, the error rate of the junior staff person was unknown, and if too high and unchecked, could invalidate the study. On the other hand, resources were not available for DE.

Since the initiation of the SCS, our organization had purchased Teleforms,[16] an optical character recognition (OCR) software system for automated data entry, to support data capture in an unrelated study. Several formal investigations have been performed evaluating the error rate using paper forms originally designed in the Teleforms designer program (traditional forms [TF]) (Figure 1) and error rates have been found to be comparable to DE.[8,17–19]

Teleforms had recently released a new existing forms (EF) feature, in which forms designed in other programs could be transferred to the Teleforms system and the data could be scanned in using Teleforms' OCR capabilities. The error rates reported for TF encouraged us, but on further investigation, we realized that when data collection is initiated on paper forms designed in other programs, many of the OCR-enhancing features associated with TF cannot be used. We were not able to find any reliable Teleforms EF error rate benchmarks.

Table 1 ■ Error Rates for Existing Forms and Single Entry: First and Second Study

Comparison	First Study		Second Study	
	Existing Forms*	Single Entry*	Existing Forms*	Single Entry*
Overall	270	36	839	16
Circle choice	1,551	0	5,393	0
Box choice	146	44	357	0
Evaluating populated field as blank	211	8	679	4
Reading data incorrectly	59	28	160	12

*Errors per 10,000 fields.

Our research group was still interested in trying the EF feature to enter the data from the completed SCS forms, and assumed that this could be accomplished with the time resources available. But before committing to using EF method for all data entry in this ongoing study, we conducted a reliability study by comparing the error rates of EF and SE.

Methods

A set of completed SCS questionnaires was entered using EF by a data manager (M.M.W.), and the same questionnaires were entered via SE into an MS Excel[12] spreadsheet by the junior staff member. The datasets were then compared using SAS Proc Compare.[13] Any discrepancy between the values in two corresponding fields in each dataset was considered an error. For each error, the original paper form was inspected, and the correct entry was identified. Error rates per 10,000 fields were calculated in the following manner: (Number of fields in error divided by all the fields in the analysis) multiplied by 10,000. McNemar's test[20] was performed to determine whether there were significant differences in the error rates between EF and SE.

Example

We first analyzed questionnaires from 93 participants composed of 442 forms containing 17,146 fields. Overall error rates for EF (in comparison to those using SE) were surprisingly high (270 per 10,000 fields compared with 36 per 10,000

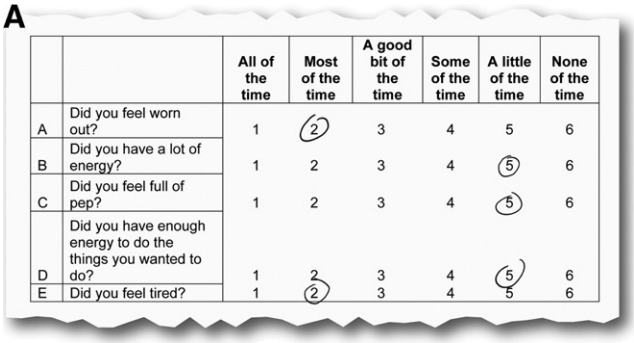

Figure 2. Examples of circle and box choice fields on forms.

fields for SE, McNemar χ^2 306.87, $p < 0.0001$) (Table 1). The EF error rate was highest in circle choice fields (1,551 per 10,000 fields), yet the error rate for box choice fields was still prohibitively high (146 per 10,000 fields) (Figure 2). Interestingly, most of EF's errors were in evaluating a field as blank rather than populated (error rate 219 per 10,000) rather than reading the data incorrectly (error rate 59 per 10,000 fields). In any case, given that most of the questionnaire fields were either circle or box choice fields, EF was not performing adequately for this data entry task (when compared with SE) in the current environment.

The Teleforms system uses an OCR engine called the reader. The system allows users to fine tune this OCR reader's performance. Initially, Teleforms options were left at manufacturer's settings. The company was contacted and adjustments were made aimed at improving EF error rates given our specific issues. We then replicated our study methods on questionnaires from 14 participants representing 70 forms containing 2,660 fields.

Unfortunately, these adjustments did not help, and the EF error rate actually increased to 839 per 10,000 fields overall, with the circle choice error rate increasing to 5,393 per 10,000 fields, and the box choice error rate to 357 per 10,000 fields. EF now read populated fields as blank at an error rate of 679 per 10,000 fields, and incorrectly read values from fields at an error rate of 160 per 10,000 fields.

Given these results, we decided to carry out data entry in the rest of the study using SE only, as the SE error rate of 36 per 10,000 fields was comparable to the literature and acceptable to us.

Discussion

The Teleforms EF system used in the study did not achieve an error rate that compared favorably with published SE, DE, or Teleforms TF error rates. However, in the process of determining this, we learned the error rate of SE for our junior staff member, and found it to be acceptable. The results of this study provided us the necessary information to select SE as our data entry method for the remainder of the study.

Many efforts focused on transcribing health data to an electronic format do not evaluate the data quality of the electronic result. Harding et al.[21] report converting to an interactive voice response (IVR) system for data collection to replace their paper system in their research study. Although they do not conduct a comparison of data quality, they conclude that IVR "... represents a marked advantage in ... data collection in large multicentre trials ...."[21] Likewise, Puskar et al.[22] defend choosing Teleforms TF for their multicenter study, claiming "... this software product allows for ... more accurate data entry ..."[22] without conducting data quality checks.

Formal data quality studies often start after data collection has begun, and deem whatever error rate they find acceptable. Quan et al.[23] gathered reliability data in their palliative care research study, reporting overall rates of TF recording incorrect data and missing data at 0.4% and 0.6% of 980 data elements, translating to error rates of 41 and 61 per 10,000 fields, respectively. Shiffman et al.[24] aimed to reduce the incidence of missing medical record data by using Teleforms TF for data capture into an electronic format rather using paper charts, and although they succeeded in improving data completeness, a study of data accuracy was never reported.

Choosing between an automated entry system such as OCR or IVR vs. a low-tech system such as DE or SE is also a cost–benefit issue, so what constitutes an acceptable error rate can vary widely depending on a project's goals. An error rate of 100 or more per 10,000 fields may be acceptable if the cost savings are high (for example, one study with an error rate this high calculated cost savings of $1,900.08 per questionnaire entered[3]). A limitation of our current study is that, unlike other studies of Teleforms and other automated entry products,[3,8,18,19,23,25] it did not collect cost–benefit metrics. On the other hand, if an acceptable error rate is selected *a priori* at the time of conception of the data collection effort, this can serve as a guide for the choice of a data entry system.

From our experience, we would recommend a best-practice paradigm of planning for the transcription of data from paper forms to electronic format during the forms design/selection stage, rather than afterward. Although our study compared the error rate of a unique application of a specific technology with the error rate of a particular individual's data entry performance, we feel this general approach would increase the success of data entry projects under other circumstances. Advance planning may involve a literature review to arrive at a consensus on an acceptable error rate given the costs, pilot testing of data entry to ensure feasibility and to calculate actual error rates, and collaboration with data management experts. Unfortunately, these efforts may not always be possible. If data collection has commenced using paper forms without a data entry system set up, the "next to the best practice" is to pilot test a data entry system and conduct a formal evaluation before committing to it for the entire duration of the project. This approach will help to promote informed, defensible, and cost-effective decisions about which data entry system to use during research studies that start with paper forms.

References ■

1. Neaton JD, Duchene AG, Svendsen KH, Wentworth D. An examination of the efficiency of some quality assurance methods commonly employed in clinical trials. Stat Med 1990;9:115–23, discussion 24.
2. Hakansson I, Lundstrom M, Stenevi U, Ehinger B. Data reliability and structure in the Swedish National Cataract Register. Acta Ophthalmol Scand 2001;79:518–23.
3. Smyth ET, McIlvenny G, Barr JG, Dickson LM, Thompson IM. Automated entry of hospital infection surveillance data. Infect Control Hosp Epidemiol 1997;18:486–91.
4. Microsoft Office Word. 2003 ed. Redmond, WA: Microsoft Corporation, 2006.
5. Microsoft Office Access. 2003 ed. Redmond, WA: Microsoft Corporation, 2006.
6. Oracle 11g. 11g ed. Redwood Shores, CA: Oracle Corporation, 2007.
7. Microsoft SQL Server. 2005 ed. Redmond, WA: Microsoft Corporation, 2006.
8. Jorgensen CK, Karlsmose B. Validation of automated forms processing. A comparison of Teleform with manual data entry. Comput Biol Med 1998;28:659–67.

9. Weber BA, Yarandi H, Rowe MA, Weber JP. A comparison study: paper-based versus web-based data collection and management. Appl Nurs Res 2005;18:182–5.
10. Gibson D, Harvey AJ, Everett V, Parmar MK. Is double data entry necessary? The CHART trials. CHART Steering Committee. Continuous, hyperfractionated, accelerated radiotherapy. Control Clin Trials 1994;15:482–8.
11. Reynolds-Haertle RA, McBride R. Single vs. double data entry in CAST. Control Clin Trials 1992;13:487–94.
12. Microsoft Office Excel. 2003 ed. Redmond, WA: Microsoft Corporation, 2006.
13. Statistical Analysis Software. 9.1 ed. Cary, NC: SAS Institute, 2006.
14. Goldin SB, Wahi MM, Wiegand LR, et al. Perspectives of third-year medical students toward their surgical clerkship and a surgical career. J Surg Res 2007;142:7–12.
15. Goldin SB, Wahi MM, Farooq OS, et al. Student quality-of-life declines during third-year surgical clerkship. J Surg Res 2007; 143:151–7.
16. Teleforms (Cardiff). 9.1 ed. San Francisco, CA: Verity, 2006.
17. Dyck PJ, Turner DW, Davies JL, O'Brien PC, Dyck PJ, Rask CA. Electronic case-report forms of symptoms and impairments of peripheral neuropathy. Can J Neurol Sci 2002;29:258–66.
18. Guerette P, Robinson B, Moran WP, et al. Teleform scannable data entry: an efficient method to update a community-based medical record? Community care coordination network Database Group. Proc Annu Symp Comput Appl Med Care 1995:86–90.
19. Jinks C, Jordan K, Croft P. Evaluation of a computer-assisted data entry procedure (including Teleform) for large-scale mailed surveys. Comput Biol Med 2003;33:425–37.
20. Swinscow TD. Statistics at square one. XVI-The chi squared 2 tests. Br Med J 1976;2:573–4.
21. Harding JP, Hamm LR, Ehsanullah RS, et al. Use of a novel electronic data collection system in multicenter studies of irritable bowel syndrome. Aliment Pharmacol Ther 1997;11: 1073–6.
22. Puskar KR, Lamb J, Boneysteele G, Sereika S, Rohay J, Tusaie-Mumford K. High touch meets high tech. Distance mental health screening for rural youth using Teleform. Comput Nurs 1996;14:323–9; quiz 30–2.
23. Quan KH, Vigano A, Fainsinger RL. Evaluation of a data collection tool (TELEform) for palliative care research. J Palliat Med 2003;6:401–8.
24. Shiffman RN, Brandt CA, Freeman BG. Transition to a computer-based record using scannable, structured encounter forms. Arch Pediatr Adolesc Med 1997;151:1247–53.
25. Nies MA, Hein L. Teleform: a blessing or burden? Public Health Nurs 2000;17:143–5.