Can Prospective Usability Evaluation Predict Data Errors?

Constance M. Johnson, PhD, RN[1], Meredith Nahm, PhD[2], Ryan J. Shaw, MS, RN[1] , Ashley Dunham, PhD[2], Kristin Newby, MD[1], Rowena Dolor, MD[1], Michelle Smerek[2], Guilherme Del Fiol, MD, PhD[1], Jiajie Zhang, PhD[3]
[1]Duke University, Durham, NC [2]Duke Translational Medicine Institute, Durham, NC
[3]University of Texas at Houston, School of Biomedical Informatics, Houston, TX

Abstract

Increasing amounts of clinical research data are collected by manual data entry into electronic source systems and directly from research subjects. For this manual entered source data, common methods of data cleaning such as post-entry identification and resolution of discrepancies and double data entry are not feasible. However data accuracy rates achieved without these mechanisms may be higher than desired for a particular research use. We evaluated a heuristic usability method for utility as a tool to independently and prospectively identify data collection form questions associated with data errors. The method evaluated had a promising sensitivity of 64% and a specificity of 67%. The method was used as described in the literature for usability with no further adaptations or specialization for predicting data errors. We conclude that usability evaluation methodology should be further investigated for use in data quality assurance.

Introduction

A recent literature review demonstrated that single data entry errors in clinical research averaged 78 errors per 10,000 fields and ranged from 3.8 – 650[1]. Today, data entry errors are of growing concern because more data are collected via patient entry, *e.g.*, electronic patient reported outcomes, or otherwise recorded initially in computer data systems. When the initial recording is data entry into a computer, there is no opportunity to go back afterwards and correct errors, *i.e.*, there is no source for comparison and we must depend on prevention. Previous research demonstrates that there is a relationship between system usability and medical error and data quality [2, 3].

To err is human, even in systems with exemplar design. Nevertheless, data entry systems that are designed to match user requirements and expectations, thus modeled on the tasks of the users result in the mitigation of error[4]. Systems require not only sound utility but also usability, to decrease user error, thus improving data quality, and to optimize user productivity and satisfaction. However, often due to time and cost constraints or even unfamiliarity with user-centered design, many system designers fail to address usability principles in the design of data entry systems[5].

There is one easy usability inspection method that can prospectively uncover problems with a user interface, indicate the severity of the problems and make suggestions for fixing both global and local problems. Heuristic evaluation (HE) is one of the most commonly used inspection techniques due to its low cost[6]. It is a method that can uncover both major and minor problems not necessarily found with user testing[7]. Although major problems are generally easier to discover than minor problems and are the most important to fix, minor problems can just as easily contribute to data entry errors and are easier to find via HE than by other evaluation methods[7]. Furthermore, HE often picks up minor usability problems not found with user testing[7]. Although generally used prior to user testing, it can be used alone if the reviewers are double experts, trained in the domain being evaluated and in usability testing. HEs performed by two or more usability experts can identify >50% of the usability problems with an interface[7]. While there are other usability related design and evaluation methods such as cognitive task analysis and small-scale usability studies, due to time and cost constraints, HE is a sound selection because it focuses on multiple dimensions of system usability issues.

Excellent reviews of methods applied to research data collection, specifically computer aided interviewing exist[8]. However, use of usability methods to build predictive models of user performance or data quality in research data collection has not been reported. Further, empirical comparisons of the data quality from different data collection mechanisms, i.e. drop down boxes versus radio buttons have been reported[9]. However, these evaluations were performed outside of a usability construct.

In contrast, we employed a HE of a research data collection system in parallel with a data quality

assessment of the same system. Our objective was to measure the ability of HE to prospectively identify form questions more likely to produce lower data quality.

Usability testing methods, including HE, are used within an iterative design process for computer system interfaces. Such additional evaluation and testing adds time to system development. In clinical research, development of systems often occurs as configuration of vended products, and within timeframes measured in weeks rather than months. Possibly for these reasons, prospective or design phase usability testing is not commonly employed within clinical research. Our institution does not routinely employ usability evaluation as part of the design process for clinical research data collection systems. Further, usability evaluation is not mentioned in the good clinical data management practices document (GCDMP), a popular best practices document employed in clinical research data management[10].

Usability testing matches the user's mental model with the designers' conceptual model[11]. As such, HE may be effective in predicting data error, and ultimately in improving data quality. In clinical research where increasingly, data are initially collected electronically at the point of care or directly from patients[12], cleaning data as a post collection activity is problematic, as there is no source for comparison. Thus, the user interface at the initial data collection becomes the sole opportunity to increase data quality. In this context, prospective methodology to evaluate data collection interfaces and identify fields likely to produce data errors would benefit researchers.

Methods

This empirical observational study compares usability evaluation and data accuracy assessments of a clinical research data collection system. The evaluated system was customized for the Measurement to Understand Reclassification of Disease of Cabarrus/Kannapolis (MURDOCK) study from an existing system.

The MURDOCK study is a 50,000 participant community registry conducted in Cabarrus County, North Carolina, by a collaboration of Duke Researchers and leaders from the local community. The registry is part of a larger research effort to better understand chronic diseases by using new genetic and molecular tools in combination with clinical information.

A Web-based application was developed to support the various activities related to the MURDOCK study. The first version of the application was released in January 2009 to support the enrollment of study participants. New modules are under development, including 1) recruitment, study management, and participant contact management queues; 2) data importation of public listings to maximize recruitment; 3) Web and kiosk-based forms for self-enrollment; 4) follow-up forms; 5) laboratory specimen tracking via data exchange interfaces; 6) HL7 interfaces with electronic health record systems to retrieve and store participant clinical data; and 7) online reports. The application is housed in a secure, HIPAA-compliant environment. To facilitate recruitment in various locations, members of the study office staff can access the application from any computer with access to the Internet.

Usability Evaluation: HE was performed independently by two evaluators. Each evaluator assessed the online data entry screens for violations of 14 different usability heuristics. The 14 usability heuristics included visibility (users should always be informed of the system state), consistency (interface design standards and conventions should be employed), match (user model matches system model), minimalist (limited use of extraneous information), memory (minimize memory load such as recognition v. recall), flexibility (shortcuts to accelerate performance), message (good error messages), error (prevent errors), closure (clear closure on all tasks), reversible actions (undo functions), control (avoid surprising actions), feedback (provide informative feedback about actions), language (use users' language), and document (help and documentation)[7, 11].

Two independent evaluators generated separate lists of heuristic violations. These lists were compiled into one comprehensive list. Each evaluator independently scored the severity of the problems on a scale from 1, indicating a cosmetic problem (fix can wait), to 4, indicating a catastrophic problem (immediate fix)[7]. As a guideline for scoring, we considered the proportion of users who would experience the problem and whether the problem would be persistently encountered or whether it would be a problem only once. Persistent problems with a major impact received the highest rating. These scores were averaged and a final violation list with averaged scores was compiled. If there were significant differences between the scores, such as cosmetic versus catastrophic error, the evaluators discussed the differences and the scores were

updated to reflect the outcome of the discussion. Finally, solutions to the violations were provided.

Data accuracy evaluation: The data accuracy evaluation was conducted according to the database quality control standard operating procedure employed by the Duke Clinical Research Institute. The evaluation is a "standard data collection form to database" audit as described in the Good Clinical Data Management Practices document. The database audit was conducted in May/June 2009.

A sample size of 42 cases, with 7000 total fields, was planned to provide a one sided 95% confidence interval of 20 errors per 10,000 fields. The sample size was based on a field count of 166 fields for the enrollment, lab collection and medication forms. These forms represent data collected at enrollment in the MURDOCK study. This was the initial audit for the study, conducted within the first six months of enrollment at which time, there were 760 participants enrolled. The 42 case samples were selected randomly. Data listings were printed from the database. These data listings were compared to the paper data collection forms at the study office on the research campus in Kannapolis, NC.

During the audit all discrepancies between the data listings and data collection forms were recorded. In the audit closing meeting, the discrepancies were reviewed with the study data manager to confirm that the discrepancy represented an error, and were assigned a root cause. The findings were summarized in the final audit report dated July 14, 2009 for the data team and study leadership.

Comparison: The usability evaluation and data audit were conducted and reported in complete isolation. Data were combined for this analysis in October 2009. Both the usability evaluation and the database audit indexed findings by form question. For each form question, the number of data errors was counted. Also, for each form question, the presence of a usability finding was noted. Typographic errors from the usability analysis were not counted because the impact on data entry errors was considered minimal.

Results

Usability Results: A total of 56 usability problems were identified from the 14 usability heuristics earlier described earlier. 39% (n = 22) were cosmetic, 38% (n = 21) were categorized as minor, 18% (n = 10) were major, and 5% (n = 3) were catastrophic.

"Error" (n = 26) was the most frequently violated usability heuristic. Fourteen of these violations were minor usability problems such as the study only allows participants from one state, yet all 50 states are shown in the drop-down list. Ten of the "Error" violations were major usability problems such as not setting parameters on fields (i.e. allowing birth dates in the 1700's). Two problems were categorized as catastrophic. One of the catastrophic problems allowed users to enter partial telephone numbers but did not give a warning message that these telephone numbers were not saved if only partial data were entered.

"Consistency" (22 identified problems) was the second most frequently violated heuristic. Nineteen of the violations were cosmetic such as misspellings of words or inappropriately bolded fields. Three of the problems were categorized as minor such as misalignment of fields or entire questions.

The third most violated heuristic was "Minimalist". Of the 4 violations, 3 were cosmetic such as displaying the word, "Commands" above the list of command buttons (which is obvious that they are commands) and 1 minor problem where there was not enough white space between questions.

There were additionally two "Match" violations. Of the two "Match" problems, one was cosmetic and the other minor. For example, under the contact information the best time to call was not listed in logical order (p.m. followed by a.m.).

Finally, there was one major "Flexibility" violation where subjects had to use the calendar to enter date of birth, and there was one catastrophic "Feedback" violation. The "Feedback" violation involved providing the users with a list of data entry errors at the bottom of the screen after they saved their entire entry. The bottom of the screen was not always clearly visible if the user was at the top of the screen.

Database audit results: The database error rate was 75 errors in 6804 fields audited, normalized to 110 errors per 10,000 fields with a 95% confidence interval of (87,139), consistent with the published literature for single entered data1. The 75 errors were categorized by error type (Table 1).

Error type	Number of errors	Number of fields audited*	Error Rate**
Data entry	40	6804	59
Interpretations on Medications	15	840	179
Process	10	6804	15
Check box / number consistency	7	84	833
system limit	2	Na	Na
Lab label	1	42	238

** number of fields audited reflects the number of fields subject to the error type
* error rates normalized to errors per 10,000 fields

Table 1. Data Error Categorization

A total of 40 errors were attributed to data entry and 33 to process (compliance with written work instructions for data collection or entry), and 2 to human interaction with fields that did not contain parameters. All 75 errors were used in the comparison with the usability evaluation. The process-related data errors were included because it is possible that system usability can impact user ability to follow established procedures.

Comparison results: Findings from the HE were overlaid on the database errors (Figure 1). In Figure 1 each box represents a question on the data form. Shaded boxes correspond to usability findings. The numbers are counts of database errors occurring on a given form question. Shaded boxes with numbers show usability finding. Eighteen form questions had coincident data error and usability findings, 10 had database errors only, and 27 had usability findings only, 25 questions had neither. Thus, the HE had a sensitivity of 64% and a specificity of 67%.

1	**1**		**8**		**1**	**5**	
1		**3**				**2**	
	2						
3					**3**		
	6	**1**		**1**			
	3	**1**	**1**				
	3						**2**
	1	**1**		**1**	**1**	**2**	**1**
	1					**1**	**18**

Figure 1. Coincidence Usability Finding - Data Error Occurrence

In the case of the question with eight data errors, the usability evaluation identified the root cause of the error, such as the system not saving incomplete telephone numbers. This root cause was not determined in the audit. The usability analysis identified seven other fields with three or more errors. It is likely that the usability issues represented underlying systematic causes of the data errors that were not otherwise obvious.

Discussion

HE provides an orthogonal and systematic method to assess data systems. While other methods such as cognitive task analysis, which examines aspects of design that may impose user cognitive burden, and small-scale usability studies, which may detect some problems not found with HE, add to identifying usability problems within an interface, we raised the question of what one test would identify the majority of usability problems in this cost-constraining climate. Therefore, we made a trade-off and considered that HE would reveal the majority of problems, since it focuses on multiple dimensions and not just errors.

HE revealed both major and minor usability problems. "Error" was the most frequently violated usability heuristic. We found that both major and minor "error" problems could be solved with simply inserting field parameters. "Consistency" is one of the most violated usability problems[7] and was the second most violated heuristic in our analysis. Although the majority of the consistency violations were scored as cosmetic, issues with consistency can confuse users if they have to wonder whether different words or actions mean the same thing. For example simple issues with spatial consistency such as misalignment of fields can easily induce data entry errors. The third most violated heuristic was "Minimalist", which as Nielsen states, "less is more."[7] Extraneous information tends to distract users and slows down data input. Problems such as not including enough white space between questions can make text difficult to read and has the potential to not only slow users down, but can also incite errors. Another heuristic violated was that of "Flexibility", which gives users the flexibility of creating customizations and shortcuts to accelerate their performance. Although using a calendar for a date field helps users easily identify dates, using it for a date of birth field just adds time and frustration to data entry for both novice and experienced users. The users should be allowed to enter the dates manually.

While there were other minor, easily fixed heuristic violations, one of the most catastrophic violations occurred with the "Feedback" violation where users were provided a list of problems at the bottom of the screen after saving their entries. Users should always

be given prompt and informative feedback at the point of data entry. This type of violation requires the users to go back and find the errors taking unnecessary time for data entry. It is possible that the users may not even see these violations at the time they save their entry.

Independent means of evaluation such as comparison to independent data sources, double data entry, and database audit are commonly used to increase data quality. Therefore, pursuing HE for this application is warranted. Although the sensitivity and specificity presented herein are on the low side of predictability, our results show that HE has a promising sensitivity and specificity that warrants further investigation and methodological evaluation. Importantly, form questions with higher numbers of errors tended to also to have usability findings such as feedback on valid value violations after the conclusion of the data entry process. Questions with higher numbers of usability errors are most likely to be associated with systematic rather than random errors. Our results suggest that HE has utility in indicating fields subject to systematic error. In data collection, where data quality practices such as post-entry data cleaning, double data entry, or validation with independent data sources are not feasible, few tools are left for practitioners to increase data quality; mainly on-screen data checks (missing, range, valid value) and internal data consistency checks. There is a need in clinical research data management to use methods based on studies in human-computer interaction, to assess data entry interfaces that could inadvertently introduce data entry errors. In this case, HE can be used prospectively to identify fields that may be prone to systematic error, thus HE may provide an additional mechanism for increasing data quality.

Conclusion

This analysis indicates that HE has potential utility in data quality. Based on the results here, application of HE as one type of rapid and inexpensive test for data quality should be investigated. Since this study is basically a feasibility study, future research in this area needs to concentrate on the specificity and sensitivity of this testing. Further, cost benefit analysis of such testing in clinical research data collection will inform decision makers on the benefits of this test.

Acknowledgements

This project was supported by the MURDOCK study, and by Grant Number UL1RR024128 from the National Center for Research Resources (NCRR), and its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH.

References

1. Nahm M, Johnson C, Johnson T, Fendt K, Zhang J. Clinical research data quality literature review and systematic analysis. Submitted for review (Clinical Trials Journal). 2009.
2. Obradovich JH, Woods DD. Users as designers: how people cope with poor HCI design in computer-based medical devices. Human Factors. 1996;38(4):574-92.
3. Strong DM, Lee YW, Wang RY. Data quality in context. Communications of the ACM. 1997;40(5):103-10.
4. Johnson C, Johnson T, Zhang J. Increasing productivity and reducing errors through usability analysis: a case study and recommendations. Proceedings - AMIA Symposium. 2000:394-8.
5. Food and Drug Administration: Center for Devices and Radiological Health. Human factors implications of the new GMP rule. Overall requirements of the new quality system regulations. Journal [serial on the Internet]. April 1998 Date: Available from: http://fda.gov/chrh/humfac/hufacimp.html.
6. Cockton G, Lavery D, Woolrych A. Inspection-based evaluations. In: Jacko JA, Sears A, editors. The human-computer interaction handbook Fundamentals, evolving, technologies and emerging applications. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers; 2003.
7. Nielsen J. Usability engineering. Boston: Academic Press; 1993.
8. Couper MP. Usability evaluation of computer-assisted survey instruments. Social Science Computer Review. 2000(18):384.
9. Heerwegh D, Loosveldt G. An evaluation of the effect of response formats on data quality in web surveys. Social Science Computer Review. 2002;20(4):471-84.
10. Society for Clinical Data Management. Good clinical data management practices. Journal [serial on the Internet]. 2009 Date: Available from: http://www.scdm.org.
11. Shneiderman B. Designing the user interface. Strategies for effective human-computer interaction. Reading, MA: Addison Wesley Longman, Inc; 1998.
12. Schuyl ML, Engel T. A review of source documentation verification process in clinical trials. Drug Information Journal. 1999;33:789-97.