Spring 5-6-2010

Data Accuracy in Medical Record Abstraction

Meredith Nahm
University of Texas Health Science Center at Houston

Dissertation

Data Accuracy in Medical Record Abstraction

By

Meredith Nahm, MS

May 6, 2010

APPROVED:

Jiajie Zhang, PhD

Todd Johnson, PhD

Constance M. Johnson, PhD, RN

Jack Smith, MD, PhD

Amy Franklin, PhD

Data Accuracy in Medical Record Abstraction

A

DISSERTATION

Presented to the Faculty of
The University of Texas
School of Health Information Sciences
at Houston
in Partial Fulfillment
of the Requirements

for the Degree of

Doctor of Philosophy

by

Meredith Nahm, MS

Committee Members:

Jiajie Zhang, PhD[1]
Todd Johnson, PhD[1]
Constance M. Johnson, PhD, RN[2]
Jack Smith, MD, PhD[1]
Amy Franklin, PhD[1]

[1]University of Texas, School of Health Information Sciences, Houston TX, [2]Duke University School of Nursing, Durham, NC

Preface

Medical record abstraction is the process in which a human manually searches through a medical record to identify data required for a secondary use. Abstraction involves some direct matching of information found in the record to the data required, but also includes operations on the data such as categorizing, coding, transforming, interpreting, summarizing, and calculating. The abstraction process results in a summary of information about a patient for a specific secondary data use. Medical Record Abstraction remains a primary mode of data collection in clinical research, quality improvement, performance measurement, disease surveillance, and other secondary data uses.

While hundreds of articles mention factors that may impact the accuracy of abstracted data, the information in the literature until now has not been synthesized, and the majority of the work has been done in the absence of a theoretical framework.

Information generation, collection, and representation are central to informatics. Generation, collection, and representation impact data and information quality; in turn, data and information quality impact use. Medical Record Abstraction is about the interaction of humans with the processes, tools, representations, and environment in which the abstraction occurs. In medical record abstraction, a human being is an agent in the collection and transformation of data. That human-data-representation interaction is an informatics problem that, until now, has not yet been addressed from an informatics perspective.

The work presented here was motivated by the lack of consensus and lack of evidence supporting methods used in the collection and management of clinical research data and their impact on the quality of the data. This work began with a quantitative literature review and pooled analysis of data error rates reported in the clinical trial and registry literature. This first paper included in this compilation associated medical record abstraction with the highest error rates of the data collection and processing methods common in clinical research. Thus, data quality in medical record abstraction became the focus of further investigation.

The second paper in the compilation, a formal concept analysis of data quality, was necessary for further investigation of data quality in medical record abstraction. The concept analysis clarified the multidimensionality of data quality, and focused my work on the dimension of data accuracy, *i.e.,* correctness of the data values.

My study of data accuracy in medical record abstraction was initiated with a review and formal synthesis of the medical record abstraction literature. Working with the literature helped identify appropriate theoretical frameworks and led the way to a classification system for factors impacting the accuracy of medical record abstraction. The factors impacting data accuracy in medical record abstraction reported in the literature were assed (content validity) through a two cohort, four round Delphi process. The third paper in this compilation presents the results of this work.

The fourth and final paper in this compilation investigates one factor, cognitive load, consistently indicated in the literature as impacting data accuracy in medical record abstraction. Representational Analysis methodology was applied to asses the possibility that abstractor cognitive load during abstraction reaches published limits of human cognition. This work demonstrated that cognitive load during abstraction from characteristics of the data elements alone, not only reaches, but in 9% of the data elements, exceeds human cognitive limits.

This work lays the groundwork for additional research and furthers both the science of informatics and the clinical and translational research to which it is applied.

Dedication

To Oscar David Biddy, who taught me that work should be fun, and
to Deborah Roth , Rob Califf, and my colleagues at Duke who made it so.

"Blessed is he who has found his work …"
−*Thomas Carlyle, Past and Present, 1843*

To William Van Orden Nahm, who taught me that anything was possible;
To Rosemary Andrews, who believed that I could achieve anything;
To Rosemary Nahm, who, by example showed me how;
To William Ed Hammond and Jiajie Zhang, who guided me along the way;
To Leonard White who joined me on the journey; and with whom I look forward to many more.

Acknowledgements

I would like to acknowledge my dissertation chairman, Jiajie Zhang, PhD, and my doctoral committee: Todd Johnson, PhD, Constance M. Johnson, PhD, RN, Jack Smith, MD, PhD, and Amy Franklin, PhD for their guidance during my doctoral program and research. I would also like to acknowledge the faculty, staff and students in the University of Texas Health Science Center at Houston, School of Health Information Sciences (SHIS) to whom I will always be grateful for what I have learned during my years in the SHIS program, especially, to Jim Turley, RN, PhD for feedback on and input into the Delphi pilot, and Annath Annapragada, PhD for his very helpful feedback during the writing of my dissertation proposal.

I would also like to acknowledge the following for their assistance in operationalizing my doctoral research: Rosemary A. Nahm, RN, MSN, who served as the independent interviewer for the Delphi and the independent reviewer for the literature synthesis; Vickie Nguyen, MA and Elie Razzouk, MD for their independent review for the representational analysis; the Duke Translational Medicine Institute UL1 RR024128-01 programming support for the Delphi Round 1 and 2 surveys, and the administrative support of Ashley Talley for getting many, many articles and scheduling Delphi interviews; and the Society for Clinical Research Associates (SoCRA) and the American Health Information Management Association (AHIMA) for allowing me to recruit research participants at their annual conferences.

Without these organizations and individuals, this research would not have been possible.

Table of Contents

CLINICAL TRIALS

Clinical Research Data Quality Literature Review and Pooled Analysis

Journal:	**Clinical *T*rials*: J*ournal o*f* the *S*ociety *f*or Clinical *T*rials**
Manuscript ID:	draft
Manuscript Type:	Review
Date Submitted by the Author:	
Complete List of Authors:	Nahm, Meredith; Duke Translational Medicine Institute, Clinical Data Integration Johnson, Constance; Duke University, School of Nursing Johnson, Todd; University of Texas - Houston, School of Health Information Sciences Fendt, Kaye; Duke University Medical Center, Duke Clinical Research Institute Zhang, Jiajie; University of Texas - Houston, School of Health Information Sciences
Keywords:	Data management, Data quality, Clinical research, Registries, Electronic data processing, Chart review, Medical record abstraction, Data collection

Abstract:	We present a literature review and secondary analysis of data accuracy in clinical research and related secondary data uses. A total of 93 papers meeting our inclusion criteria were categorized according to the data processing methods. Quantitative data accuracy information was abstracted from the articles and pooled. Our analysis demonstrates that the accuracy associated with data processing methods varies widely, with error rates ranging from 2 errors per 10,000 files to 5019 errors per 10,000 fields. Medical record abstraction was associated with the highest error rates (70–5019 errors per 10,000 fields). Data entered and processed at healthcare facilities had comparable error rates to data processed at central data processing centers. Error rates for data processed with single entry in the presence of on-screen checks were comparable to double entered data. While data processing and cleaning methods may explain a significant amount of the variability in data accuracy, additional factors not resolvable here likely exist.

Clinical Research Data Quality Literature Review and Pooled Analysis

Meredith Leigh Nahm, MS[a,d], Constance Johnson, PhD[c], Todd Johnson, PhD[d], Kaye Fendt, MSPH[b] and Jiajie Zhang, PhD[d]

From the [a]Duke Translational Medicine Institute, Durham, NC, [b]Duke Clinical Research Institute, Durham, NC, [c]Duke University School of Nursing, Durham, NC, and the [d]School of Health Information Sciences, University of Texas, Houston, TX

Word count: 6766

Corresponding author:

Meredith Nahm, MS

Duke Translational Medicine Institute

5162 Hock Plaza, 2424 Erwin Road

Durham, NC 27705

Tel: 919-668-8339

E-mail: meredith.nahm@duke.edu

Abstract

We present a literature review and secondary analysis of data accuracy in clinical research and related secondary data uses. A total of 93 papers meeting our inclusion criteria were categorized according to the data processing methods. Quantitative data accuracy information was abstracted from the articles and pooled. Our analysis demonstrates that the accuracy associated with data processing methods varies widely, with error rates ranging from 2 errors per 10,000 files to 5019 errors per 10,000 fields. Medical record abstraction was associated with the highest error rates (70–5019 errors per 10,000 fields). Data entered and processed at healthcare facilities had comparable error rates to data processed at central data processing centers. Error rates for data processed with single entry in the presence of on-screen checks were comparable to double entered data. While data processing and cleaning methods may explain a significant amount of the variability in data accuracy, additional factors not resolvable here likely exist.

Abstract Word Count: 150 words

Key Words: Data management; Data quality; Clinical research; Registries; Electronic data processing; Chart review; Medical record abstraction; Data collection

Introduction

Computers have been used in clinical research since the early 1960s, although initial efforts to integrate them into such efforts were experimental and sporadic [1,2]. The early application of computers to health-related research spawned a plethora of methods for collecting and preparing data for analysis [3,4]. These functions, which are central to the field of clinical research informatics, are evaluated by metrics of cost, time, and quality [5]. While cost and time affect the feasibility of research, and timeliness is certainly critical to the conduct and oversight of research, the scientific validity of research conclusions depends on data quality in general and data accuracy in particular [6].

Reports on data accuracy assessments begin in the late 1970s [7] and continued throughout the 1980s [8-12]. Although data accuracy is likely attributable to how data are collected, entered, and cleaned, or otherwise processed, no systematic reviews of the clinical research data quality literature have been conducted to date. In 2002, Arts published a review of data quality in medical research registries [13]. Although a significant contribution, Arts' work was limited in scope to registries and did not include other secondary uses, such as clinical trials.

Several texts describe approaches to general data quality management independent of the domain area of the application [14-17]. These works focus on general methods for assessing and documenting data quality, as well as methods for storing data in ways that maintain or improve their quality. As such, the works abstract above data collection and processing methods applicable to specific industries and types of data. They provide little to no guidance to investigators and research teams planning a clinical research endeavor or attempting to operationalize data collection and management.

Background

Many authors in the clinical research arena lament the paucity of published information regarding clinical trials data quality [5,6,8,12,18-26]. While most authors point out that conclusions drawn from studies depend on data quality (and the underlying data collection and management methods), others label the associated tasks mundane [27,28]. This perception has perhaps resulted in the minimal degree of investigation and associated low number of publications on the topic of data collection and management, compared with other areas of clinical research and informatics methodology. An initial step in seeking such information and providing guidance to investigators and research teams is a thorough review and synthesis of the relevant published literature.

With the current rapid influx of new technology into healthcare and clinical research, the quality of data quality achievable using available methods becomes even more important; e.g., learning from the past, and providing benchmarks for new technology evaluation [29]. Many unresolved issues exist with respect to data quality in clinical research, including a thorough understanding of the accuracy and variability of current data processing methods. Thus, a synthesis of information about clinical research data quality from the literature is a critical contribution to both clinical research informatics and clinical and translational research.

Historically, data processing methods used across clinical research show significant variability [3,4,30]. Briefly, according to Medical Subject Headings (MeSH) terminology, registries are the ‘‘systems and processes involved in the establishment, support, management, and operation of registers (sic), e.g., disease registers.” In a 2008 paper, Drolet distinguishes registries from simple databases or other non-registry data repositories and presents six criteria for *medical data registries* [31]. Not evident in his definition, however, are the wide variety of

data processing methods used for medical data registries. These include: 1) chart review and abstraction versus direct electronic acquisition from electronic medical records; 2) use of vended data collection systems by local healthcare facilities (e.g., data entry and cleaning in local systems versus Web-based data entry and cleaning in a centrally hosted system); 3) use of paper data collection forms and central processing; and 4) single versus double data entry. Data cleaning methods also vary greatly, from use of "reports of things to consider for the next data submission," to on-screen checks during data entry, to post-entry batch data processing.

Except for the ongoing transition from paper-based data collection to Web-based collection, methods used to gather and process data for industry-sponsored clinical trials performed in support of marketing authorization have remained fairly uniform over the 30-year span of the literature base. Generally, in this highly regulated variety of clinical research, data are abstracted from patient charts, then transcribed onto either paper forms or electronic data capture (EDC) systems. Where paper forms are used, data are forwarded to a central data center, double-entered, and subjected to data cleaning checks. Discrepancy reports are sent to the clinical investigational sites, and resolutions to any queries are received in return; based on these responses, the database is then updated. In contrast, data collected on modern EDC trials are generally single-entered at clinical investigational sites in centrally hosted Web-based systems that employ on-screen data cleaning checks.

Data collection and processing for clinical research funded by non-industry means (such as trials funded by governmental and foundational resources), varies widely. As with registries and market-oriented clinical research, the process typically starts with chart review, also called *medical record abstraction* (or simply "abstraction"). Small early-phase or low-budget projects,

however, often employ commercially-available spreadsheets for data collection and management, and manual data review for data cleaning.

Unfortunately, while use of spreadsheets in research is ubiquitous, there are few reports on how this affects data quality [32-36]. Both small and mid-size projects often use Microsoft Access (Redmond, WA) or similar pseudo-relational data systems that provide utilities for creating data entry screens, a scripting language for error checks, and a data storage mechanism. Data entry in such systems tends to be single entry, with a combination of on-screen and post-entry data cleaning.

Similarly, some investigators choose to enter data directly into statistical analysis packages such as SPSS (Chicago, IL) or SAS (Cary, NC). Larger government- and foundation-funded projects tend to use custom data collection systems developed specifically for a given project; many of these use relational database management systems. There has been a recent trend in government-funded projects toward specialized, vended clinical data management systems employed in regulated clinical trials. These systems support multiple workflows, and different data processing methodologies.

Thus, historically, data collection and processing methods have varied greatly across types of clinical research, and there has been little comparative evaluation of different methods and technologies. Key events, however, signal an increasing convergence in methods. These include: 1) industry contract research organizations serving as data centers for government-funded research; 2) use of central data centers for government-funded work; 3) a recent increase in academic membership in the historically industry-based Society for Clinical Data Management; 4) National Institutes of Health (NIH) contract and grant solicitations that require compliance with industry regulations; and 5) government adoption of vended clinical data management

systems [30]. Convergence of such great variability necessitates comparative evaluation of methods and technologies of data collection and management with respect to cost and data quality [29].

Significant variability also exists in quantitative methods for assessing data accuracy across clinical research and other secondary data uses [13,37,38]. Data accuracy has often been measured in terms of database error rates (registries also often assess percent completeness, and, thus evaluate data quality in terms of data accuracy and completeness). The error rate is defined in the Good Clinical Data Management Practices document (GCDMP) as the number of errors divided by the number of data values inspected [37], a method used in other industries as well [39].

$$Error\ rate = \frac{number\ of\ errors}{number\ of\ values\ inspected}$$

As described in the GCDMP and elsewhere [40], there are significant differences in the way errors and values are inspected and counted. Based on these counting differences, the error rates obtained can differ by a factor of two or more [37,38]. In addition, differences in how error rates are reported (e.g., as raw counts, errors per record, or errors per 10,000 fields), necessitate scaling and normalization of the values reported in the literature before meaningful comparisons can be made.

Assessment and quantification of data accuracy is crucial to scientific inquiry. Errors increase variance, which in turn diminishes statistical power. Three papers report simulations of power loss or related constructs for different error rates [23,41,42]. Rostami *et al.* provide an analytical solution for the impact of database errors on reliability and show the resulting increase

in sample size necessary to compensate for the loss of statistical power for given error rates in the dependent analysis variable [40].

The effect of errors on data quality is dependent on the variable in which the error occurs; i.e., in major independent or dependent variables or covariates [43]. The impact is also dependent on the robustness of the particular analytical method used. Thus, there is no regulatory or one-size-fits-all standard error rate for clinical research. A 1999 Institute of Medicine report emphasizes this by defining quality data in clinical research as "data that support the same conclusions as error free data [6]."

One balancing factor, the cost and time associated with cleaning data, can account for upwards of 10%–20% of the cost of a clinical trial [44]. Thus, cleaning data to appropriate quality levels and not beyond is clearly the target. To support this, a thorough characterization of data processing methods with respect to data accuracy is clearly needed to identify the most cost-effective balance. To that end, we undertook a review and secondary analysis of the relevant literature to characterize data processing methods with respect to quality.

Purpose

We sought to synthesize the literature on data quality in clinical research and registries. Four main questions comprise the focus of this research:

1. How much, and what kind of, quantitative data quality information is reported in the literature?
2. What data collection and processing methods are described? Secondarily, is the information applicable to newer data processing methods (such as Web-based EDC) in clinical research?

3. Do error rates differ by data processing method? If so, how and by how much?
4. Are some data processing methods inherently more variable than others? If so, to what extent?

Methods

Criteria for inclusion of manuscripts

The criteria for inclusion in this analysis were as follows: 1) articles must be manuscripts published in peer review journals indexed for retrieval, be referenced by such, or be publicly available; 2) have a focus on secondary data use (e.g., clinical research, quality improvement, surveillance, research registries); 3) the database error rate must be presented or resolvable (e.g., via number of errors identified and number of fields inspected, or must contain sufficient information to calculate); 4) must describe how the data were processed (e.g., optical scanning, single or double entry); 5) must be written in the English language; and 6) must be a primary source for the error rate. The following parameters were optional: data cleaning method, location of data processing (central data center vs. local healthcare facility), gold standard used, and scope of method of comparison.

Information retrieval

A PubMed search on the Medical Subject Heading (MeSH) terms *"data quality" AND (registry or "clinical trial")* produced 350 citations through 2008. Review of the 350 abstracts produced 54 articles deemed likely to contain quantitative information about data quality; the full text of these 54 articles was reviewed. PubMed related links and secondary and tertiary references from

the 54 articles identified an additional 70 articles of interest. After full-text review of the 124 articles, 93 articles met our criteria and were used in this analysis.

The literature

The literature in this area spans publication dates from 1978–2008. The clinical research data quality literature base is fragmented across 60 distinct journals, including medical specialty journals, statistics journals, clinical research journals, and informatics journals. Journals with more than one article meeting our inclusion criteria are listed in Table 1.

Although articles could be categorized by the clinical area of interest, categorization by type of secondary use and data processing method is more germane to our investigation. The literature is categorized into major groupings according to type of secondary use, data processing method, and data accuracy assessment method in Table 2. The categories in Table 2 represent combinations of common factors and are not mutually exclusive; thus, some articles appear in more than one category. Importantly, all articles used in our analysis are included in Table 2.

Data collection and analysis

Three types of data were collected from each article: 1) information about how data were processed [Table 3]; 2) information about how data quality was measured [Table 4]; and 3) number of errors and number of fields inspected. Prior to quantitative data analysis, the factors shown in Tables 3 and 4 were developed in a qualitative, iterative manner during the review of the articles. Attributes of the data processing and quality measurement as reported were noted as each article was read. Natural groupings were organized into the categories displayed in Tables 3 and 4. These categories

were later explored in the analysis to ascertain which (if any) of the factors might affect data quality.

In the literature, data quality is presented in several different ways, including overall agreement, *kappa,* (κ), number of erroneous fields per number of records reviewed, number of erroneous fields per number of fields (values) reviewed, number of errors per number of keystrokes, and specificity and sensitivity. Traditionally, where a gold standard is used for comparison, overall agreement is reported, whereas if the researcher believes there is no gold standard (often the case of two independent raters), κ is reported. We abstracted the number of errors reported and the total number of data fields (values) inspected. In one case, an author was contacted to confirm the number of fields inspected. Four articles [9,41,88,102] presented only normalized error rates as errors per 10,000 fields; for these, the denominator was assumed for the total number of fields inspected. The number of errors and number of fields inspected were used to calculate normalized error rates (number of errors per 10,000 fields) based on the recommendations in GCDMP [37].

Each article described a data quality assessment of one or more databases. Where error rates for more than one database or assessment were provided in an article, each is included in this analysis. Likewise, in some articles, error rates were reported for more than one process step; for example, medical record to Case Report Form (CRF), CRF to first entry, first entry to second entry, or CRF to clean file. Where error rates for multiple data processing steps were provided, we include them here; thus, the number of data points is higher than the number of reviewed papers. A total of 22 articles reported results for more than one processing step, process or database [9,11,13,22,32,54,55,63,64,66,74,79-81,83-85,87,96,102,114,121], providing a total of 125 data points.

For consistency, one rater was used to abstract the error rate information from the articles (MN). A sample of the manuscripts included in the analysis, comprising 10% of the total, was re-evaluated following the initial abstraction to assess reliability. For the sample, the time between the initial and intra-rater reliability review was at least 1 year. Intra-rater reliability, calculated as percent difference, was used to gauge reliability of the data. The intra-rater reliability for number of errors, number of fields, and error rate were 85%, 97%, and 86% respectively. In addition, a second rater (CJ) reviewed the same intra-rater reliability sample, with comparable results. In light of the underlying variability in the data, the variability in error rate calculation methods currently in use, and the aims of this study, these were considered reasonable. In addition, they are comparable to those in a similar review paper of errors in electronic medical records [123].

Two reviews of data accuracy in electronic medical records report variability in the methods used to measure accuracy [123,124]. Both reviews concluded that the methodological variability was sufficiently significant to preclude pooled analysis. However, Hogan and Wagner stated that assessment methods in the field of clinical research were more uniform, an assessment similar to our own. We therefore pursued a pooled analysis and normalized error rates from the literature. From this analysis, we present descriptive statistics and confidence intervals.

Results

The data obtained from our literature review are displayed in Figure 1, which shows all 125 data points normalized as number of errors per 10,000 fields and demonstrates the dispersion over time of the health-related research literature with respect to data accuracy. Database error rates ranged from 2–5019 errors per 10,000 fields. This three orders-of-magnitude range necessitated a logarithmic display. There appears to be no pattern in the year-to-year reporting. It is reasonable

to assume that if adoption and implementation of new technology tended to increase data quality, such a trend would be evident in the graph. One possible explanation for the lack of such a trend is that adoption and implementation of new technology throughout the years has not increased the data accuracy. Another possibility is that data accuracy is not considered in new technology evaluation and implementation.

In order to explore possible differences in data quality for different data processing methods, data were categorized according to the method used to process the data, shown in Figure 2. Source-to-database and source-to-CRF were combined into a single category labeled *abstraction* based on: 1) some of the articles reported error rates for abstraction directly to an electronic data collection form; i.e., no separate data entry step, and 2) the central tendency and dispersion of the two processes being similar. Data have been horizontally jittered to more fully display the data set. The median (horizontal bar) is overlaid on each group. The source-to-database error rates ranged from 82–5019 errors per 10,000 fields; the source-to-CRF error rates ranged from 70–3360 errors per 10,000 fields. Optical scanning error rates ranged from 2–1106 errors per 10,000 fields. Single-entry error rates ranged from 4–650 errors per 10,000 fields, while double-entry error rates ranged from 4–33 errors per 10,000 fields. Additional descriptive statistics are provided in Table 5.

Several salient features are apparent in Figure 2. First, ordered by the median, medical record abstraction is associated with the highest error rate, followed by optical methods and single entry, while double data entry, the least complex and most controlled, is associated with the lowest error rate. Importantly, abstraction, optical methods, and single entry all are associated with significant variability. Notably, the error rates reported for both medical record abstraction

and optical methods span three orders of magnitude. For optical methods, there were too few data points to support a subgroup analysis.

For single entry, there were both a sufficient number of data points and detail regarding different methods to perform a subgroup analysis (Figure 3), in which we examined the effects of aspects of data processing reported in the literature, including the presence of on-screen checks, use of batch data cleaning, and the location of data processing. These variations are examined here because they represent key data processing options in current Web-based EDC systems, but as they constitute small subgroups, any findings from their analysis should be considered only as indicators of potential trends. For example, labels such as “batch data cleaning” or “on-screen checks” were only applied where the original manuscript specifically so noted; therefore, the other data points may or may not have employed batch data cleaning or on-screen checks. Location of data processing (e.g., at local facilities versus central data centers), was discernable from all articles.

Analysis of the subgroup data revealed the following patterns. Location of data processing appeared to have little effect on data accuracy. When both distributed and centrally entered data are considered together, on-screen checks (OSC) were associated with a decrease in average error rates, from 158 errors per 10,000 fields to 23; in addition, variability was correspondingly decreased from 211 errors per 10,000 fields to 15 (the categories labeled “All” in Figure 3). The trend is the same for data entered at clinical sites (“distributed OSC” in Figure 3), where on-screen checks were associated with a decreased average error rate of 28 errors per 10,000 fields, comparable to the average quality obtained from double entry. Importantly, the data presented in Figure 3 represent only data processing accuracy—they do not include errors that may result

from upstream processes (such as errors in the medical record itself or errors arising from abstraction process). Descriptive statistics for Figure 3 are shown in Table 6a.

The predominant method of clinical research data collection in the near future will likely be Web-based EDC. Due to the importance of this particular model, articles reporting data accuracy from similar data processing configurations (e.g., central versus distributed data processing in the presence of OSC), were examined (Figure 4). Similar to the single-entry subgroup results, this analysis showed little difference in error rates for centrally versus locally processed data. Thus, the EDC model is likely capable of providing accuracy comparable to that of centrally processed data. As in Figure 3, the data presented in Figure 4 represent only data processing accuracy. Descriptive statistics are provided in Table 6b.

Finally, we examined the effect of batch data cleaning on error rates. This analysis was performed on the entire data set. Only 19 of the studies we examined reported on whether batch data cleaning techniques were employed. We conservatively grouped articles that explicitly stated that no batch data cleaning was used with articles that made no mention of batch data cleaning. This analysis (Figure 5) showed that batch data cleaning was associated with lower error rates and decreased variability (descriptive statistics are reported in Table 7). Databases employing batch data cleaning were more likely to use central data processing and double data entry. In addition, the upper mode of the batch-cleaned data, with the exception of one data point, consisted of medical record abstraction studies.

Discussion

Over the last 30 years, empirical data accuracy assessments have been reported in the literature. Although there have been articles that summarized results reported in multiple papers

[5,13,21,22], there has been little synthesis, and this area of inquiry still lacks a theoretical framework. Further, although notable exceptions exist [11,22,33,59,79,84,85,87,92], little evidence has been obtained regarding the relative accuracy of different data processing methods. Characterizing the factors that affect data accuracy, as well as the differences in accuracy among popular methods (as done inductively here), serves as the beginning of such a framework.

Medical record abstraction

First and foremost, our results indicate that medical record abstraction is associated with error rates an order of magnitude greater than that attributable to "downstream" data processing techniques. These results support claims that medical record abstraction is the most significant source of error in clinical research [21,125]. This is unfortunate, because medical record abstraction remains the dominant method of data collection in retrospective and prospective research, as well as in the fields of safety surveillance and healthcare quality improvement.

Although position papers and reports of empirical results exist, the reasons for high error rates associated with medical record abstraction have not been systematically studied, and the mechanisms are not understood. Empirical studies suggest that there is significant variability in the abstraction and quality control processes used [126,127]; these different methods, process aids and quality control activities could be responsible for the amount of variation observed. Likewise, the reliance on human performance and associated underlying cognitive processes could be responsible for some or all of the variability.

Several authors have further explored these underlying reasons for the high variability in abstraction [56,81,125-129]. Differences related to therapeutic area, such as type of data collected, patient acuity, and chart complexity, are one source of variability [81,129].

Additionally, variables recorded at multiple time points and at multiple places in the patient chart, as well as those that require the abstractor to interpret or synthesize information into a score, are known to be more subject to error [81], as are variables requiring the use of clinical judgment on the part of the abstractor [128]. Further, although many investigators proceed on the premise that charts provide correct data [130], there are a significant number of errors, inconsistencies, and omissions in the charts themselves [126,130]. Such chart errors, while not arising from abstraction, nonetheless further increase the variability in abstracted data, and are not measured or accounted for by current data quality assessment practices [129]. Exploring the causes of this variability is an important area for future research. Unfortunately, while there are sufficient data points, there is insufficient information in the literature about the medical record abstraction methods to further investigate possible causes for the variability in a subgroup analysis.

Importantly, medical record abstraction errors are less likely to be detected by downstream data processing such as data entry or programmatic data cleaning. For example, an incorrect but plausible value chosen from the medical record will not be detected by valid range checks. Medical record abstraction errors that result in plausible values will only be detected through comparison to the medical record; i.e., re-abstraction, while errors in the medical record itself will likely remain undetected by this method.

Although differences in medical charts among sites complicate the construction of abstraction guidelines for multicenter research, differences between abstractors are so significant that most authors have recommended and employed detailed abstraction guidelines for each variable [56,81,125-128]. Abstraction guidelines are used in national performance measurement and healthcare quality improvement programs, such as the Joint Commission specifications for

performance measures [131,132] and the Healthcare Effectiveness Data and Information Set (HEDIS) roadmap [133].

However, such guidelines, which constitute a primary mechanism for preventing abstraction errors, are not often used in clinical trials. Instead, less specific form-completion instructions are employed. These instructions often do not specify from where in the medical record a particular value should be obtained, or provide guidance in the case of multiple available or missing values. In a process called *source data verification*, data for a sample of subjects are compared to the chart, typically without calculating an error rate [81,134]. It is unfortunate that the association of clinical trials with prospective data collection further fuels the perception that abstraction or chart review is not a factor in data accuracy when in fact the chart is the source of most clinical research data, and abstraction from the chart is the part of the data collection process most subject to error [21,125]. Despite recommendations for measuring and monitoring data quality from the medical record abstraction process [29,127,128], abstraction error usually remains unquantified [29,40,81,134].

Optical methods

Although methods such as optical character recognition (OCR) and optical mark recognition (OMR) have been touted as a faster, higher-quality or less resource-intensive substitute for manual data entry [27,58,62,84,92,93,135-137], others have reported error rates with optical methods that were three times higher than manual keyboard data entry [138]. As was the case with medical record abstraction, optical methods were associated with a variability of three orders of magnitude in accuracy, which is clearly unacceptable for a data entry method. Such variability may be influenced by: the presence and type of data cleaning employed in processing

the optical scans; use of post-entry visual verification or pre-entry manual review; training of form completers on handwriting; differences in form compatibility with the software; software configuration (e.g., recognition engine); and variations in data quality assessment methods. For example, some implementations employ programmatic checks and visual verification of fields labeled with low confidence from the OCR or OMR engine. In addition, the different engines employed need to be calibrated for each implementation.

Reports on the accuracy of optical methods uniformly showed greatest accuracy with mark objects (check boxes, fill-in-bubbles), followed by numeric fields [27,92,138]. The accuracy of free-text fields was the lowest of the three field types [58,92,138]. Some authors have reported comparisons of optical methods with double or single data entry [32,62,92,138]; however, because of the variability in the reported accuracy, together with the relative paucity of reports and the lack of sufficient detail regarding associated data processing methods, we did not summarize these findings separately. The results here suggest that with appropriate adjunct data cleaning processes, optical methods can achieve comparable accuracy to single-entered data. The wide variability also suggests that an *a priori* quantitative assessment of data accuracy and total cost is necessary when an optical system is newly installed or upgraded.

Keyboard entry

A review of the literature provides the opportunity for analyzing different variations on key entry, including single entry, single entry with on-screen data checks, and double entry. We undertook these comparisons because single entry is widely used with Web-based EDC, the extent of OSCs is at the discretion of the study team, and there is little evidence to guide practice. Our analysis demonstrates that single entry with OSC is associated with accuracy comparable to

that obtained with double data entry. However, single entry alone was associated with an average of more than 100 errors per 10,000 fields, compared with an average of 23 errors per 10,000 fields with OSC. The subgroup analysis of single entry provides a plausible explanation for the variability. This is an important finding, because large amounts of data are collected from research sites via Web-based systems, including entry of abstracted data into Web-based systems, clinicians entering data in electronic health records, and data collected directly from patients via hand-held devices. Due to the problem of "alert fatigue," however, OSC may not be feasible in electronic medical records, where clinical alerts will often be a higher priority. The question of alert fatigue in these systems is an important topic for further research.

Central versus distributed

Given the movement toward Web-based EDC and the use of electronic medical record data in clinical research, the question of the relative accuracy between data entered centrally versus distributed data entry is an important one. Our results show that in the presence of OSCs, data entered centrally have marginally better accuracy compared with data entered at investigational sites. However, with the exception of one outlier, the points for distributed data entry fell within the range of the centrally entered data. Given the variability in implementations of OSC (e.g., number and type of checks), these results suggest that data quality equivalent to that of centrally-entered data can be obtained through distributed methods, and that more information is needed to guide practice.

Batch data cleaning

Batch data cleaning was associated with a decrease in error rates of more than 300 errors per 10,000 fields, as well as an almost 50% decrease in variability. In addition, an internal analysis at our institution noted that on average, values for 1%–2% of any variable changed based on batch data cleaning. Although this number was based on programmed batch checks and double entered data, the number is consistent with the results from the literature review, and likely to be considerably higher for single-entered data. However, as noted earlier, use of batch data cleaning was reported in only 19 of the cases reviewed in this analysis (manuscripts that did not report batch cleaning were conservatively assumed not to have used it). Importantly, batch data cleaning is resource-intensive to implement, because it requires programming logic for each suspected discrepancy, the number of which are exponentially combinatorial with the number of data fields collected. In addition, the asynchronous identification of data discrepancies (i.e., after data entry) necessitated by batch data cleaning also adds a costly feedback cycle, often involving multiple iterations, with the clinical investigational site. Batch data cleaning also suffers from variability in implementation, including number and type of checks programmed, type of data checked, and number of iterations with sites. For these reasons, registries and quality improvement reporting do not routinely use batch data cleaning, and the costs of batch data cleaning (cited at $35–$100 per discrepancy for paper-based trials [139-141]), has led to their utility being questioned in the context of multicenter clinical trials [141].

Relative accuracy differences between methods

The amount of overlap in the quality levels across data processing methods suggests that different data processing methods or combinations thereof can achieve comparable accuracy.

Data cleaning methods that are applied during or after data entry, while creating additional costs and user burdens, are likely to increase the accuracy obtainable from a data entry method. For example, use of OSCs was associated with lower error rates and lower variability, whether data were entered at clinical sites or at a central data center. Additionally, distributed data entry with OSCs was associated with error rates similar to those obtained from double-entered data. This could be attributed to the proximity to source information where data are entered at clinical investigational sites. The latter two comparisons are critical, because adoption of Web-based EDC necessitates single data entry at clinical investigational sites. The information presented here should prompt those using EDC to fully utilize OSCs.

Perception of acceptability regarding different error rates varies from "10% is good" to "1% is good" to "there should be no errors [6,20,38]." The answer to the question, "How clean is clean enough?" is multifaceted. The effect of errors is dependent on the variable in which the errors occur, the number and extent of the errors, and the robustness of the analytical method [43]. Based on simulations showing an approximate reduction in statistical power of 1% for each additional percent of the error rate [23,41,42] and the analytical work of Rostami *et al* [40], the abstraction error rates seen in the published literature are high enough to impact upon the decisions made using the data. Abstraction is a crucial method of data collection, and will remain so until data in the healthcare setting are sufficiently standardized to support automated extraction from EMRs. Until that time, data will continue to be manually extracted from both paper and electronic records. The association of abstraction with high error rates and significant variability, as noted here, supports recommendations that both abstraction methods and measures of accuracy should be reported with the results from studies employing medical record abstraction [126,127].

Limitations

This study is a secondary pooled analysis of database error rates in the published literature. Although it constitutes a critical contribution in synthesizing a very fragmented literature base, there are many limitations inherent in our work. First, there is the possibility that we may have missed relevant articles. Second, because our work is a secondary analysis, it relies on data that were collected for other purposes. Although we used error and field counts reported in the literature, prior work has shown that even these have significant variability [37,38]. For example, some may count dates as discrepant if there is not an exact match, while others may allow a window of several days; field counts may exclude null fields, or include fields entered once and propagated to multiple places. Results presented here should be examined in this context. Third, a lack of standard terminology for data processing methods potentially affects this analysis through high likelihood that relevant manuscripts were not identified as mentioned above, and also through misinterpreting the information presented in the source manuscripts. Fourth, due to the small number of data points in some subgroups, comparisons between data cleaning methods are less reliable than for the four higher-level categories of data processing methods. Fifth, some of the articles we included in our analysis reported an assessment of only a few variables, while others analyzed many variables or an entire database. Given the observed variability in data quality for different therapeutic areas, likely driven by the differences in the type of data, reports of error rates on only 1 or a limited number of variables are less generalizable. Sixth, most of the articles in our review were from academic organizations and government or foundation-funded endeavors that employ different data collection and management methodologies. Although over

the time span of the literature we reviewed, those methods have tended to converge, our results may be less applicable to industry funded studies.

Given these limitations, we draw conclusions only about associations and trends. Importantly, although there is strong correlation between data processing method and both central tendency and dispersion of data error rates, this association does not imply causality, and other important factors not assessable here may be responsible for these results. Other possible explanations for the variability seen here include variations in local system implementations and workflows, the number of data processing steps, differences in reporting error rates, skill level of the staff, and complexity of the data. These factors were not assessable in a literature review such as ours. The inability to measure or control for these possible alternatives represents a weakness in this analysis, and they remain topics for further investigation.

Further research

The published information on data quality derives mainly from evaluation and observation of data accuracy in funded registries and clinical trials, as opposed to targeted experimental inquiry. While a biannual conference on cross-industry data quality exists [142], and a clinical research-focused Data Quality Research Institute was founded in 2004 [143], funding for dedicated data quality research has remained slim or non-existent; therefore, there are few controlled experiments evaluating data capture or cleaning methods to date and even less generalizable work contributing to the body of knowledge. Research from multiple epistemological stances would provide valuable information to confirm or challenge the trends identified here. Further, as data (increasingly captured electronically) are used to support direct patient care, performance measurement, and research, the effects of data quality on decision-making need thorough

exploration, as do the effects of system usability and data entry and cleaning methods on data quality and clinical workflow.

Conclusions

The formal literature on clinical research data quality remains significantly fragmented. Standardizing terminology and using it to index the literature base may lead to wider recognition of the applicability of data quality work within clinical research and biomedical informatics and provide amalgamation points. Importantly, the results presented here, while gained from a secondary analysis, are of immediate use in informing investigators and research teams of the options and characterization of common data collection and processing methods with respect to data accuracy. In particular, the variability if not the magnitude of error rates reported should encourage evaluation of the impact of new technology and processes on data accuracy, and subsequent decisions regarding whether the accuracy is acceptable for the data's intended use.

In synthesis, we suggest that data quality varies widely by data processing method. Further, it appears that the lowest-quality process, abstraction, is the most ubiquitous and also the least measured and controlled within research projects. Although they pale in comparison to abstraction error rates as a source of variability, the differences in accuracy of data entry methods also deserve consideration. On average, double data entry provides the highest accuracy and lowest variability, followed by single entry. Of all popular data entry techniques, optical methods provide the lowest data quality and the highest variability. However, other confounding factors assessed, including use of OSC with single data entry, and batch data cleaning checks, as well as factors not identifiable or assessable in a secondary analysis (e.g., staff experience, number of

manual steps, and data complexity), may constitute substantial mediators with the potential to equalize differences between methods.

Error rates reported in the literature are well within ranges that could necessitate increases in sample sizes from 20% or more in order to preserve statistical power for a given study design [144,145]. Data errors have also been shown to change p values [44] and attenuate correlation coefficients to the null hypothesis [42,146,147]; in other words, a given clinical trial may fail to reject the null hypothesis because of data errors rather than because of a genuine lack of effect for the experimental therapy [149]. In the presence of large data error rates, a researcher must then choose to either 1) accept unquantifiable loss of statistical power and risk failure to reject the null hypothesis due to data error; or 2) measure the error rate and increase the sample size to maintain the original desired power [40,145,147]. The adverse impact of data errors has also been demonstrated in registries and performance measurements [121,149-154], as has failure to report data [155].

While such results in aggregate are shocking, we do not present them to incite panic or cast doubt upon clinical research results. Other factors that are not assessable here, such as variables in which the errors occurred, and statistical methods used to take the measurement error into account, are necessary for such assessments. We applaud the authors of the reviewed papers for their rigor and forthrightness in assessing error; measurement is the first step in management. We hope that our analysis makes a strong and convincing argument for the measurement and publication of data accuracy in clinical research.

While there is general agreement that the validity of research rests on a foundation of data, data collection and processing are sometimes perceived as menial and peripheral to the core operations of caring for patients or completing a clinical research project. In between rote data

entry and scientific validity, however, lie many unanswered questions, which, if answered, will help investigators and research teams balance cost, time and quality while assuring scientific validity.

Acknowledgements

We gratefully acknowledge the following people, many of them authors on source manuscripts included in this review, who freely shared raw data or information about how data in their studies were processed: Marie Diener-West, PhD and Marvin Newhouse for providing data from their 1998 paper on data quality in the Collaborative Ocular Melanoma Study; David Christiansen, PhD for correspondence about the ARIC study activities; Eva Dreisler, MD and Sven Adamsen, MD for information on how the Laparoscopic Cholecystectomy Registry data were processed; Chung-Yi Li, PhD for information on how data in the Taiwan Birth Registry were processed; Fu-Lin Wang, PhD for information on how data were processed in the Alberta Birth Registry; Rosemary Cress, DrPH for information about how the colorectal cancer data in the California Cancer Registry were processed; Torin Block of Nutritionquest for assistance locating the CARDIA dietary history form. Without their correspondence, the data set created here would have lacked several important works. We also gratefully acknowledge Jonathan McCall for editing versions of this manuscript, and Eric Reyes for creating the figures.

This work was supported in part by the Clinical and Translational Science Awards (CTSA) to Duke University (1 UL1 RR024128) and to the University of Texas Health Science Center at Houston (5 UL1 RR024148)

References

1. **Forrest WH Jr, Bellville JW**. The use of computers in clinical trials. Br J Anaesth 1967; **39**: 311–319.
2. **Helms RW**. Data quality issues in electronic data capture. Drug Inf J 2001; **35**: 827–837.
3. **Helms RW**. A distributed flat file strategy for managing research data. The ACM Portal Web site. http://delivery.acm.org/10.1145/810000/809982/p279-helms.pdf?key1=809982&key2=2993683421&coll=GUIDE&dl=GUIDE&CFID=38484650&CFTOKEN=70124220. Accessed May 29, 2009.
4. **Collen MF**. Clinical research databases--a historical review. J Med Syst 1990; **14**: 323–344.
5. **Knatterud GL, Rockhold FW, George SL, et al**. Guidelines for quality assurance in multicenter trials: a position paper. Control Clin Trials 1998; **19**: 477–493.
6. Division of Health Sciences Policy, Institute of Medicine. *Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making: Workshop Report: Roundtable on Research and Development of Drugs, Biologics, and Medical Devices*. Washington, DC: National Academy Press; 1999.
7. **Kronmal RA, Davis K, Fisher LD, Jones RA, Gillespie MJ.** Data management for a large collaborative clinical trial (CASS: Coronary Artery Surgery Study). Comput Biomed Res 1978; **11**: 553–566.
8. **Bagniewska A, Black D, Molvig K, et al.** Data quality in a distributed data processing system: the SHEP Pilot Study. Control Clin Trials 1986; **7**: 27–37.
9. **Prud'homme GJ, Canner PL, Cutler JA**. Quality assurance and monitoring in the Hypertension Prevention Trial. Hypertension Prevention Trial Research Group. Control Clin Trials 1989; **10**: 84S–94S.
10. **DuChene AG, Hultgren DH, Neaton JD, et al**. Forms control and error detection procedures used at the Coordinating Center of the Multiple Risk Factor Intervention Trial (MRFIT). Control Clin Trials 1986; **7**: 34S–45S.
11. **Crombie IK, Irving JM**. An investigation of data entry methods with a personal computer. Comput Biomed Res 1986; **19**: 543–550.
12. **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 1987; **8**: 96–100.

13. **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 2002; **9**: 600–611.
14. **Redman TC.** *Data Quality for the Information Age*. Norwood, MA: Artech House; 1996.
15. **Redman TC.** *Data Quality: The Field Guide*. Boston, MA: Digital Press; 2001.
16. **Batini C, Scannapieco M.** *Data Quality: Concepts, Methodologies and Techniques*. New York, NY: Springer; 2006.
17. **Lee YW, Pipino LL, Funk JD, Wang RY.** *Journey to Data Quality*. Cambridge, MA: MIT Press; 2006.
18. **Knatterud GL.** Methods of quality control and of continuous audit procedures for controlled clinical trials. Control Clin Trials 1981; **1**: 327–332.
19. **Freeland B.** Moving toward continuous quality improvement. J Intraven Nurs 1992; **15**: 278–282.
20. **Christian MC, McCabe MS, Korn EL, Abrams JS, Kaplan RS, Friedman MA.** The National Cancer Institute audit of the National Surgical Adjuvant Breast and Bowel Project Protocol B-06. N Engl J Med 1995; **333**: 1469–1474.
21. **Blumenstein BA.** Verifying keyed medical research data. Stat Med 1993; **12**: 1535–1542.
22. **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**: 115–123.
23. **Day S, Fayers P, Harvey D.** Double data entry: what value, what price? Control Clin Trials 1998; **19**: 15–24.
24. **Steward WP, Vantongelen K, Verweij J, Thomas D, Van Oosterom AT**. Chemotherapy administration and data collection in an EORTC collaborative group--can we trust the results? Eur J Cancer 1993; **29A**: 943–947.
25. **Van den Broeck J, Mackay M, Mpontshane N, Kany Kany Luabeya A, Chhagan M, Bennish ML.** Maintaining data integrity in a rural clinical trial. Clin Trials 2007; **4**: 572–582.
26. **Hall HI, Gerlach KA, Miller DS.** Methods of quality management: accuracy. Journal of Registry Management 2002; **29**: 72–77.
27. **Nies MA, Hein L**. Teleform: a blessing or burden? Public Health Nurs 2000; **17**:143–145.

28. **Crewson PE, Applegate KE.** Fundamentals of clinical research for radiologists: Data collection in radiology research. AJR Am J Roentgenol 2001; **177**: 755–761.
29. **Nahm ML, Pieper CF, Cunningham MM.** Quantifying data quality for clinical trials using electronic data capture. PLoS ONE 2008; **3**: e3049.
30. **Nahm M.** Clinical Data Management, Why we do what we do. Data Basics 2007; **13**.
31. **Drolet BC, Johnson KB.** Categorizing the world of registries. J Biomed Inform 2008; **41**: 1009–1020.
32. **Wahi MM, Parks DV, Skeate RC, Goldin SB**. Reducing errors from the electronic transcription of data collected on paper forms: a research data case study. J Am Med Inform Assoc 2008; **15**: 386–389.
33. **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**: 182–185.
34. **Panko RR**. What we know about spreadsheet errors. Journal of End User Computing 1998; **10**: 15–21.
35. **Panko RR.** Reports of Spreadsheet Errors in Practice. The Spreadsheet Research (SSR) Web site. http://panko.shidler.hawaii.edu/SSR/index.htm. Accessed May 27, 2009.
36. **Nahm M, Zhang J.** Operationalization of the UFuRT methodology for usability analysis in the clinical research data management domain. J Biomed Inform 2009; **42**: 327–333.
37. Society for Clinical Data Management. *Good Clinical Data Management Practices (GCDMP)* Milwaukee: WI, Society for Clinical Data Management; 2007.
38. **Nahm M, Dziem G, Fendt K, Freeman L, Masi J, Ponce Z.** Data quality survey results. Data Basics 2004; **10**: 13–19.
39. **Caballero I, Verbo E, Calero C, Piattini M.** MMPro: A Methodology based on ISO/IEC 15939 to Draw up Data Quality Measurement Processes. Presented at: 13th International Conference on Information Quality (2008).
40. **Rostami R, Nahm M, Pieper CF.** What can we learn from a decade of data quality audits: Impact of error rate on statistical reliability. Clin Trials 2009; **6**: 141-150.
41. **McEntegart DJ, Jadhav SP, Brown T, Channon EJ.** Checks of case record forms versus the database for efficacy variables when validation programs exist. Drug Inf J 1999; **33**: 177–224.

42. **Mullooly JP**. The effects of data entry error: an analysis of partial verification. Comput Biomed Res 1990; **23**: 259–267.
43. **Nahm M, Howard K.** Society for Clinical Data Management: Data Quality Webinar Parts 1 and 2 [webcast] 2008.
44. **Eisenstein EL, Lemons PW II, Tardiff BE, Schulman KA, Jolly MK, Califf RM.** Reducing the costs of phase III cardiovascular clinical trials. Am Heart J 2005; **149**: 482–488.
45. **Moro ML, Morsillo F.** Can hospital discharge diagnoses be used for surveillance of surgical-site infections? J Hosp Infect 2004; **56**: 239–241.
46. **McGovern PG, Pankow JS, Burke GL, et al.** Trends in survival of hospitalized stroke patients between 1970 and 1985. The Minnesota Heart Survey. Stroke 1993; **24**: 1640–1648.
47. **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 1999; **11:** 209–213.
48. **Hunt JP, Cherr GS, Hunter C, et al.** Accuracy of administrative data in trauma: splenic injuries as an example. J Trauma 2000; **49**: 679-686.
49. **Wynn A, Wise M, Wright MJ, et al.** Accuracy of administrative and trauma registry databases. J Trauma 2001; **51:** 464–468.
50. **Gibbs JL, Monro JL, Cunningham D, Rickards A.** Survival after surgery or therapeutic catheterisation for congenital heart disease in children in the United Kingdom: analysis of the central cardiac audit database for 2000–1. BMJ 2004; **328**: 611.
51. **Favalli G, Vermorken JB, Vantongelen K, Renard J, Van Oosterom AT, Pecorelli S.** Quality control in multicentric clinical trials. An experience of the EORTC Gynecological Cancer Cooperative Group. Eur J Cancer 2000; **36**: 1125–1133.
52. **Nagtegaal ID, Kranenbarg EK, Hermans J, van de Velde CJ, van Krieken JH.** Pathology data in the central databases of multicenter randomized trials need to be based on pathology reports and controlled by trained quality managers. J Clin Oncol 2000; **18**: 1771–1779.

53. **Vantongelen K, Rotmensz N, van der Schueren E.** Quality control of validity of data collected in clinical trials. EORTC Study Group on Data Management (SGDM). Eur J Cancer Clin Oncol 1989; **25**: 1241–1247.

54. **Horbar JD, Leahy KA.** An assessment of data quality in the Vermont-Oxford Trials Network database. Control Clin Trials 1995; **16**: 51–61.

55. **Jasperse DM, Ahmed SW.** The Mid-Atlantic Oncology Program's comparison of two data collection methods. Control Clin Trials 1989; **10**: 282–289.

56. **Reisch LM, Fosse JS, Beverly K, et al.** Training, quality assurance, and assessment of medical record abstraction in a multisite study. Am J Epidemiol 2003; **157**: 546–551.

57. **Fortmann SP, Haskell WL, Williams PT, Varady AN, Hulley SB, Farquhar JW.** Community surveillance of cardiovascular diseases in the Stanford Five-City Project. Methods and initial experience. Am J Epidemiol 1986; **123**: 656–669.

58. **O'Rourke MK, Fernandez LM, Bittel CN, Sherrill JL, Blackwell TS, Robbins DR.** Mass data massage: an automated data processing system used for NHEXAS, Arizona. National Human Exposure Assessment Survey. J Expo Anal Environ Epidemiol 1999; **9**: 471–484.

59. **Christiansen DH, Hosking JD, Dannenberg AL, Williams OD.** Computer-assisted data collection in multicenter epidemiologic research. The Atherosclerosis Risk in Communities Study. Control Clin Trials 1990; **11**: 101–115.

60. **Dennis B, Stamler J, Buzzard M, et al.** INTERMAP: the dietary data--process and quality control. J Hum Hypertens 2003; 17: 609–622.

61. **Hilner JE, McDonald A, Van Horn L, et al.** Quality control of dietary data collection in the CARDIA study. Control Clin Trials 1992; **13**: 156–169.

62. **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**: 425–437.

63. **Birkett NJ.** Computer-aided personal interviewing. A new technique for data collection in epidemiologic surveys. Am J Epidemiol 1988; **127**: 684–690.

64. **Velikova G, Wright EP, Smith AB, et al.** Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. J Clin Oncol 1999; **17**: 998–1007.

65. **Norton SL, Buchanan AV, Rossmann DL, Chakraborty R, Weiss KM.** Data entry errors in an on-line operation. Comput Biomed Res 1981; **14:** 179–198.
66. **Arts D, de Keizer N, Scheffer GJ, de Jonge E.** Quality of data collected for severity of illness scores in the Dutch National Intensive Care Evaluation (NICE) registry. Intensive Care Med 2002; **28**: 656–659.
67. **Arts DG, Bosman RJ, de Jonge E, Joore JC, de Keizer NF.** Training in data definitions improves quality of intensive care data. Crit Care 2003; **7**: 179–184.
68. **Tennis P, Bombardier C, Malcolm E, Downey W.** Validity of rheumatoid arthritis diagnoses listed in the Saskatchewan Hospital Separations Database. J Clin Epidemiol 1993; **46**: 675–683.
69. **Ghali WA, Rothwell DM, Quan H, Brant R, Tu JV.** A Canadian comparison of data sources for coronary artery bypass surgery outcome "report cards". Am Heart J 2000; **140**: 402–408.
70. **Herbert MA, Prince SL, Williams JL, Magee MJ, Mack MJ.** Are unaudited records from an outcomes registry database accurate? Ann Thorac Surg 2004; **77**: 1960–1964.
71. **Karam V, Gunson B, Roggen F, et al.** Quality control of the European Liver Transplant Registry: results of audit visits to the contributing centers. Transplantation 2003; **75**: 2167–2173.
72. **Lindblad U, Råstam L, Ranstam J, Peterson M.** Validity of register data on acute myocardial infarction and acute stroke: the Skaraborg Hypertension Project. Scand J Soc Med 1993; **21**: 3–9.
73. **Wibe A, Eriksen MT, Syse A, Myrvold HE, Søreide O**. Total mesorectal excision for rectal cancer--what can be achieved by a national audit? Colorectal Dis 2003; **5**: 471–477.
74. **McKinlay SM, Carleton RA, McKenney JL, Assaf AR.** A new approach to surveillance for acute myocardial infarction: reproducibility and cost efficiency. Int J Epidemiol 1989; **18**: 67–75.
75. **Reincke M, Petersenn S, Buchfelder M, et al.** The German Acromegaly Registry: description of the database and initial results. Exp Clin Endocrinol Diabetes 2006; **114**: 498–505.

76. **Strobl J, Enzer I, Bagust A, et al.** Using disease registries for pharmacoepidemiological research: a case study of data from a cystic fibrosis registry. Pharmacoepidemiol Drug Saf 2003; **12**: 467–473.

77. **Lash TL, Fox MP, Thwin SS, et al**. Using probabilistic corrections to account for abstractor agreement in medical record reviews. Am J Epidemiol 2007; **165**: 1454–1461.

78. **Al-Zahrani A, Baomer A, Al-Hamdan N, Mohamed G.** Completeness and validity of cancer registration in a major public referral hospital in Saudi Arabia. Ann Saudi Med 2003; **23**: 6–9.

79. **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 2003; **24**: 560–569.

80. **Fontaine P, Mendenhall TJ, Peterson K, Speedie SM.** The "Measuring Outcomes of Clinical Connectivity" (MOCC) trial: investigating data entry errors in the Electronic Primary Care Research Network (ePCRN). J Am Board Fam Med 2007; **20**: 151–159.

81. **Pan L, Fergusson D. Schweitzer I, Hebert PC.** Ensuring high accuracy of data abstracted from patient charts: the use of a standardized medical record as a training tool. J Clin Epidemiol 2005; **58**: 918–923.

82. **Steigler A, Mameghan H, Lamb D, et al.** A quality assurance audit: phase III trial of maximal androgen deprivation in prostate cancer (TROG 96.01). Australas Radiol 2000; **44**: 65–71.

83. **Lancaster S, Hallstrom A, McBride R, Morris M.** A comparison of key data entry versus fax data entry, accuracy and time [abstract]. Control Clin Trials 1995; **16**: 75–76.

84. **Diener-West M, Connor PB, Newhouse MM, Hawkins BS.** Feasibility of keying data from screen-displayed facsimile images in an ongoing trial: the collaborative ocular melanoma study. Control Clin Trials 1998; **19**: 39–49.

85. **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**: 482–488.

86. **Stone EJ, Osganian SK, McKinlay SM, et al.** Operational design and quality control in the CATCH multicenter Trial. Prev Med 1996; **25**: 384–399.

87. **Reynolds-Haertle RA, McBride R.** Single vs. double data entry in CAST. Control Clin Trials 1992; **13**: 487–494.

88. **Fabri J, Beller E**. An assessment of the value of double data entry and interactive checking [abstract]. Control Clin Trials 1993; **14**: 405.

89. **Caloto T.** Quality control and data-handling in multicentre studies: the case of the Multicentre Project for Tuberculosis Research. BMC Med Res Methodol 2001; **1**: 14.

90. **Kleinman K.** Adaptive double data entry: a probabilistic tool for choosing which forms to reenter. Control Clin Trials 2001; **22**: 2–12.

91. **Newhouse MM for the MPS Coordinating Center.** Data entry design and data quality [abstract]. Control Clin Trials 1985; **6**: 229.

92. **Jørgensen CK, Karlsmose B.** Validation of automated forms processing. A comparison of Teleform with manual data entry. Comput Biol Med 1998; **28**: 659–667.

93. **Quan KH, Vigano A, Fainsinger RL.** Evaluation of a data collection tool (TELEform) for palliative care research. J Palliat Med 2003; **6**: 401–408.

94. **Adams WG, Conners WP, Mann AM, Palfrey S.** Immunization entry at the point of service improves quality, saves time, and is well-accepted. Pediatrics 2000; **106**: 489–492.

95. **Rostami R.** Internal Audit Report, Hypericum Study Durham: Duke Clinical Research Institute; 2002.

96. **Rostami R**. Data quality in a global environment. Paper presented at: Society for Clinical Data Management Fall Conference; October 6, 2004; Toronto Canada.

97. **Thoburn KK, German RR, Lewis M, Nichols PJ, Ahmed F, Jackson-Thompson J.** Case completeness and data accuracy in the Centers for Disease Control and Prevention's National Program of Cancer Registries. Cancer 2007; **109**: 1607–1616.

98. **Michaud DS, Midthune D, Hermansen S, et al.** Comparison of Cancer Registry Case Ascertainment with SEER Estimates and Self-reporting in a Subset of the NIH-AARP Diet and Health Study. Journal of Registry Management 2005; **32**: 70–75.

99. **Clive RE, Ocwieja KM, Kamell L, et al.** A national quality improvement effort: cancer registry data. J Surg Oncol 1995; **58**: 155–161.

100. **Tingulstad S, Halvorsen T, Norstein J, Hagen B, Skjeldestad FE.** Completeness and accuracy of registration of ovarian cancer in the cancer registry of Norway. Int J Cancer 2002; **98**: 907–911.

101. **Rawson NS, Robson DL.** Concordance on the recording of cancer in the Saskatchewan Cancer Agency Registry, hospital charts and death registrations. Can J Public Health 2000; 91: 390–393.

102. **Kantonen I, Lepäntalo M, Salenius JP, et al.** Auditing a nationwide vascular registry--the 4-year Finnvasc experience. Finnvasc Study Group. Eur J Vasc Endovasc Surg 1997; **14**: 468–474.

103. **Volk T, Hahn L, Hayden R, Abel J, Puterman ML, Tyers GF.** Reliability audit of a regional cardiac surgery registry. J Thorac Cardiovasc Surg 1997; 114: 903–910.

104. **Löwel H, Lewis M, Hörmann A, Keil U.** Case finding, data quality aspects and comparability of myocardial infarction registers: results of a south German register study. J Clin Epidemiol 1991; **44**: 249–260.

105. **Lovison G, Bellini P.** Study on the accuracy of official recording of nosological codes in an Italian regional hospital registry. Methods Inf Med 1989; **28**: 142–147.

106. **Jenders RA, Dasgupta B, Mercedes D, Fries F, Stambaugh K.** Use of a hospital practice management system to provide initial data for a pediatric immunization registry. Proc AMIA Symp 1999: 286–290.

107. **Dreisler E, Schou L, Adamsen S.** Completeness and accuracy of voluntary reporting to a national case registry of laparoscopic cholecystectomy. Int J Qual Health Care 2001; **13**: 51–55.

108. **Cousley RR, Roberts-Harry D.** An audit of the Yorkshire Regional Cleft Database. J Orthod 2000; **27**: 319–322.

109. **Gissler M, Teperi J, Hemminki E, Meriläinen J.** Data quality after restructuring a national medical registry. Scand J Soc Med 1995; **23**: 75–80.

110. **Teperi J.** Multi method approach to the assessment of data quality in the Finnish Medical Birth Registry. J Epidemiol Community Health 1993; **47**: 242–247.

111. **Lin CM, Lee PC, Teng SW, Lu TH, Mao IF, Li CY.** Validation of the Taiwan Birth Registry using obstetric records. J Formos Med Assoc 2004; **103**: 297–301.

112. **Wang FL, Gabos S, Sibbald B, Lowry RB.** Completeness and accuracy of the birth registry data on congenital anomalies in Alberta, Canada. Chronic Dis Can 2001; **22**: 57–66.

113. **Cress RD, Zaslavsky AM, West DW, Wolf RE, Felter MC, Ayanian JZ.** Completeness of information on adjuvant therapies for colorectal cancer in population-based cancer registries. Med Care 2003; **41**: 1006–1012.
114. **Håkansson I, Lundström M, Stenevi U, Ehinger B.** Data reliability and structure in the Swedish National Cataract Register. Acta Ophthalmol Scand 2001; **79**: 518–523.
115. **Gissler M, Shelley J.** Quality of data on subsequent events in a routine Medical Birth Register. Med Inform Internet Med 2002; **27**: 33–38.
116. **Gissler M, Ulander VM, Hemminki E, Rasimus A.** Declining induced abortion rate in Finland: data quality of the Finnish abortion register. Int J Epidemiol 1996; **25**: 376–380.
117. **Lu JH, Hung JH, Lin FM, et al.** A quality study of a computerized medical birth registry. J Obstet Gynaecol 1995; **21**: 23–30.
118. **Roberts CD.** Data quality of the Drug Abuse Warning Network. Am J Drug Alcohol Abuse 1996; **22**: 389–401.
119. **Sprinson JE, Lawton ES, Porco TC, Flood JM, Westenhouse JL.** Assessing the validity of tuberculosis surveillance data in California. BMC Public Health 2006; **6**: 217.
120. **Olsson C, Eriksson N, Ståhle E, Thelin S.** The Swedish Heart Surgery Register: data quality for proximal thoracic aortic operations. Scand Cardiovasc J 2006; **40**: 348–353.
121. **Goldhill DR, Sumner A.** APACHE II, data accuracy and outcome prediction. Anaesthesia 1998; **53**: 937–943.
122. **Shiffman R, Brandt C, Hoffman M.** Scanned data entry for an electronic health maintenance record [abstract]. Pediatrics 1997; **100**: 445.
123. **Hogan WR, Wagner MM.** Accuracy of data in computer-based patient records. J Am Med Inform Assoc 1997; **4**: 342–355.
124. **Thiru K, Hassey A, Sullivan F.** Systematic review of scope and quality of electronic patient record data in primary care. BMJ 2003; **326**: 1070.
125. **Jansen AC, van Aalst-Cohen ES, Hutten BA, Büller HR, Kastelein JJ, Prins MH.** Guidelines were developed for data collection from medical records for use in retrospective analyses. J Clin Epidemiol 2005; **58**: 269–274.
126. **Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J.** Chart reviews in emergency medicine research: Where are the methods? Ann Emerg Med 1996; **27**: 305–308.

127. **Wu L, Ashton CM.** Chart review. A need for reappraisal. Eval Health Prof 1997; **20**: 146–163.
128. **Allison JJ, Wall TC, Spettell CM, et al.** The art and science of chart review. Jt Comm J Qual Improv 2000; **26**: 115–136.
129. **Simmons B, Bennett F, Nelson A, Luther SL.** Data abstraction: designing the tools, recruiting and training the data abstractors. SCI Nurs 2002; **19**: 22–24.
130. **Aaronson LS, Burman ME.** Use of health records in research: reliability and validity issues. Res Nurs Health 1994; **17**: 67–73.
131. Joint Commission on the Accreditation of Healthcare Organizations. ORYX. Available at: http://www.jointcommission.org/AccreditationPrograms/Hospitals/ORYX/ (accessed June 2, 2009).
132. Centers for Medicare & Medicaid Services and The Joint Commission. Specifications Manual for National Hospital Inpatient Quality Measures, version 2.5b. QualityNet Web site. http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier4&cid=1203781887871. Accessed June 1, 2009.
133. HEDIS Roadmap. The National Committee for Quality Assurance (NCQA) Web site. http://www.ncqa.org/tabid/585/Default.aspx. Accessed June 1, 2009.
134. **Schuyl ML, Engel T.** A review of the source document verification process in clinical trials. Drug Inf J 1999; **33**: 737–784.
135. **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**: 258–266.
136. **Smyth ET, Emmerson AM.** Survey of infection in hospitals: use of an automated data entry system. J Hosp Infect 1996; **34**: 87–97.
137. **Hardin JM, Woodby LL, Crawford MA, Windsor RA, Miller TM.** Data collection in a multisite project: Teleform. Public Health Nurs 2005; **22**: 366–370.
138. **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.

139. **Kush RD, Bleicher P, Kubick WR, et al.** *eClinical Trials, Planning and Implementation*. Boston, MA: CenterWatch; 2003.

140. **Redkar-Brown D.** Cost of a Query. In: Nahm M, personal communication via SCDM Discussion Forum. Durham, NC, 2007.

141. **Tardiff BE.** How clean is clean. Paper presented at: Drug Information Association Meeting 2000; Philadelphia, PA.

142. **Wang R.** International Conference on Information Quality (ICIQ). 2006; http://mitiq.mit.edu/iciq/. Accessed May 21, 2009.

143. Fendt K. Data Quality Research Institute (DQRI) Advancing the Quality of Clinical Data. 2004; http://www.dqri.org/. Accessed May 21, 2009.

144. **Freedman LS, Schatzkin A, Wax Y.** The impact of dietary measurement error on planning sample size required in a cohort study. Am J Epidemiol 1990; **132**: 1185–1195.

145. **Perkins DO, Wyatt RJ, Bartko JJ.** Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. Biol Psychiatry 2000; **47**: 762–766.

146. **Liu K.** Measurement error and its impact on partial correlation and multiple linear regression analyses. Am J Epidemiol 1988; **127**: 864–874.

147. **Stepnowsky CJ Jr, Berry C, Dimsdale JE.** The effect of measurement unreliability on sleep and respiratory variables. Sleep 2004; **27**: 990–995.

148. **Myer L, Morroni C, Link BG.** Impact of measurement error in the study of sexually transmitted infections. Sex Transm Infect 2004; **80**: 318–323.

149. **Williams SC, Watt A, Schmaltz SP, Koss RG, Loeb JM.** Assessing the reliability of standardized performance indicators. Int J Qual Health Care 2006; **18**: 246–255.

150. **Watt A, Williams S, Lee K, Robertson J, Koss RG, Loeb JM.** Keen eye on core measures. Joint Commission data quality study offers insights into data collection, abstracting processes. J AHIMA 2003; **74**: 20–25.

151. Hospital Quality Data: CMS Needs More Rigorous Methods to Ensure Reliability of Publicly Released Data. U.S. Government Accountability Office Web site. http://www.gao.gov/products/GAO-06-54. Accessed June 1, 2009.

152. **Braun BI, Kritchevsky SB, Kusek L, et al.** Comparing bloodstream infection rates: the effect of indicator specifications in the evaluation of processes and indicators in infection control (EPIC) study. Infect Control Hosp Epidemiol 2006; **27:** 14–22.

153. **Jacobs R, Goddard M, Smith PC.** How robust are hospital ranks based on composite performance measures? Med Care 2005; **43**: 1177–1184.

154. **Pagel C, Gallivan S.** Exploring potential consequences on mortality estimates of errors in clinical databases. Journal of Management Mathematics. Published online November 13, 2008. doi: 10.1093/imaman/dpn034

155. **Brien SE, Ghali WA.** Public reporting of the hospital standardized mortality ratio (HSMR): implications for the Canadian approach to safety and quality in healthcare. Open Medicine 2008; **2**: E1–4.

Figure legends

Figure 1 Chronology of the database error rate literature.

Figure 2 Data error rates for data processing methods. Error rates, medians and 95% confidence intervals are displayed with a logarithmic axis. The medical record abstraction process is associated with the largest and most variable error rate. Similarly, optical methods, such as optical mark recognition (OMR) and optical character recognition (OCR) as well as single data entry show variability that spans more than one order of magnitude.

Figure 3 Subgroup analysis for single entry. Error rates, medians and 95% confidence intervals for cases reporting single entry. The impact of location of data processing and the impact of on-screen (*i.e.,* at the point of entry) error checking were examined. Location had little effect on data accuracy, while presence of on-screen error checks was consistently associated with a decreased error rate.

Figure 4 Error rates for single-entered data in the presence of on-screen error checks. Figure 4 displays error rates, medians and 95% confidence intervals for single-entered data in the presence of on-screen error checks. This combination most closely resembles today's Web-based EDC process and shows that the EDC model is likely capable of producing data of comparable data processing accuracy to paper-based methods.

Figure 5 Subgroup analysis for batch data cleaning. Error rates, medians and the 95% confidence intervals from the sub-group analysis for batch data cleaning are displayed. As expected, batch data cleaning was associated with a lower error rate. In addition, the upper mode of the batch cleaned data consisted of medical record abstraction studies, showing that in-house batch data cleaning is most likely not capable of identifying and correcting errors created in the medical record abstraction process.

Table 1 Sources of articles

Journal	Number of Articles
Controlled Clinical Trials/Clinical Trials Journal	17
American Journal of Epidemiology	4
Journal of Clinical Epidemiology	3
Journal of the American Medical Informatics Association (AMIA) or AMIA Proceedings	3
International Journal of Epidemiology	2
Computers in Biomedical Research	2
International Journal of Quality in Healthcare	2
European Journal of Cancer	2
Journal of Clinical Oncology	2
Scandinavian Journal of the Society of Medicine	2
Journal of Trauma Infection and Critical Care	2
Pediatrics	2
Various medical specialty and informatics journals contributing one article	48
Total	**93**

Table 2 Categorization of articles used in this review

Literature Categorized by Type of Secondary Use and Data Processing Method
Reports of operational accuracy assessments to determine the usability of administrative databases for research or surveillance [45-50]
Reports of database quality control in clinical trials where gold standard was the medical record [12,15,20,24,51-56]
Reports of quality control in epidemiological, survey or observational studies where the gold standard was the medical record or taped interview [46,57,62]
Reports of instrument validation studies where the gold standard was the medical record or patient self report [63,64]
Reports of database quality control in research or quality improvement registries where gold standard was the medical record [13,50,57,65-78]
Reports of stand alone data processing studies using clinical trial-type forms and processing [11,33,64,79,80]
Reports of medical record abstraction evaluation [12,56,77,80,81]
Reports of clinical trial data processing quality that used the Case Report Form or other data processing step as the gold standard in a comparison to the trial database [7-10,22,25,41,59,79,82-96]
Reports of database quality control in surveillance registries [48,49,94,97-120]
Reports of data accuracy assessments of other health-related primary or secondary data uses [32,47,94,121,122]

Table 3 Data processing dimensions

Processing dimension	Description
Location of data processing	Whether data were entered and cleaned at a central data center or a local healthcare facility
Central coordinating center	Data center receiving and processing data from multiple sites. A healthcare facility does not qualify as a data center.
Local research site or healthcare facility	Location or institution where data were collected
Type of entry	Method used to convert data to electronic format
Single	One person enters data once
Double	Two different individuals enter data. Distinction was not made between different types of double data entry.
Optical mark/character recognition	Data on paper forms were optically scanned into electronic format
Source from which data were entered	Physical representation (if any) that the data enterer looked at during data entry
Medical record (paper chart)	Paper documentation of patient care produced and maintained by a healthcare facility
Electronic medical record	Electronic documentation of patient care produced and maintained by a healthcare facility
Paper data collection form	Structured data collection form completed by healthcare or research staff
Live interview or ePRO	Patient recall and direct electronic report
Image displayed on computer screen	Electronic representation of a data collection form rendered on a computer screen
Type of error checking*	Method, if any, used to clean data
On-screen during data entry	Real-time error checks that run during data entry
Post-entry batch computerized checks	Computerized error checks run after data entry, usually nightly
Visual verification	Manual comparison of entered data to the entry source

Table 4 Dimensions of data quality assessment

Process being measured
Source to database: process from medical record to database, including abstraction, data entry, and cleaning.
Source to data collection form: abstraction process from medical record to data collection form. Does not include data entry and cleaning.
Data collection form to database: process includes data entry and cleaning.
Single entry: only process measured is one entry of data.
Type of database assessed
Healthcare administrative database: coded data intended for use in billing.
Registry database: secondary database including phase 4 trials, surveillance programs, quality improvement and disease registries.
Trial database: secondary dataset collected for a clinical trial.
Case report form or data collection form: structured data collection form; secondary data.
Computer aided interview / Patient self report: dataset containing patient reported data via a computer.
Method of Measurement
Comparison to a "gold standard": manual comparison of two datasets from different sources or two different points in a process.
Independently re-processing data: data processed from the same source by an independent individual.
Linking and comparing data electronically: electronic comparison of two separate databases.
"Gold Standard" for comparison
Medical record (paper or electronic): paper or electronic documentation of care produced and maintained by a healthcare facility.
Independent or taped patient self-report: patient recall and direct electronic report.
Data collection form: structured data collection form completed by healthcare or research staff.
Independent database: database created from a separate source.
Data Independently entered from same source

Table 5. Descriptive statistics by data processing method

	Median	Mean	SD	Min	Max
Abstraction	647	960	1018	70	5019
Optical	81	219	337	2	1106
Single Entry	26	76	150	3.8	650
Double entry	15	16	10	3.5	33

SD = standard deviation

Table 6a Descriptive Statistics for Single Entry Subgroup Analysis

	Mean	Median	SD	Min	Max
Central	77	30	143	4	650
Distributed	86	25	175	15	550
Distributed, no on-screen checks	550	550	*	550	550
Distributed, On-screen Checks	28	24	17	15	69
All no on-screen checks	158	52	211	4	650
All with on-screen checks	23	22	15	4	69

*There was only one value for distributed single entry with no on-screen checks

SD = standard deviation

Table 6b Descriptive statistics for central versus distributed data processing in the subset of articles reporting use of on-screen checks

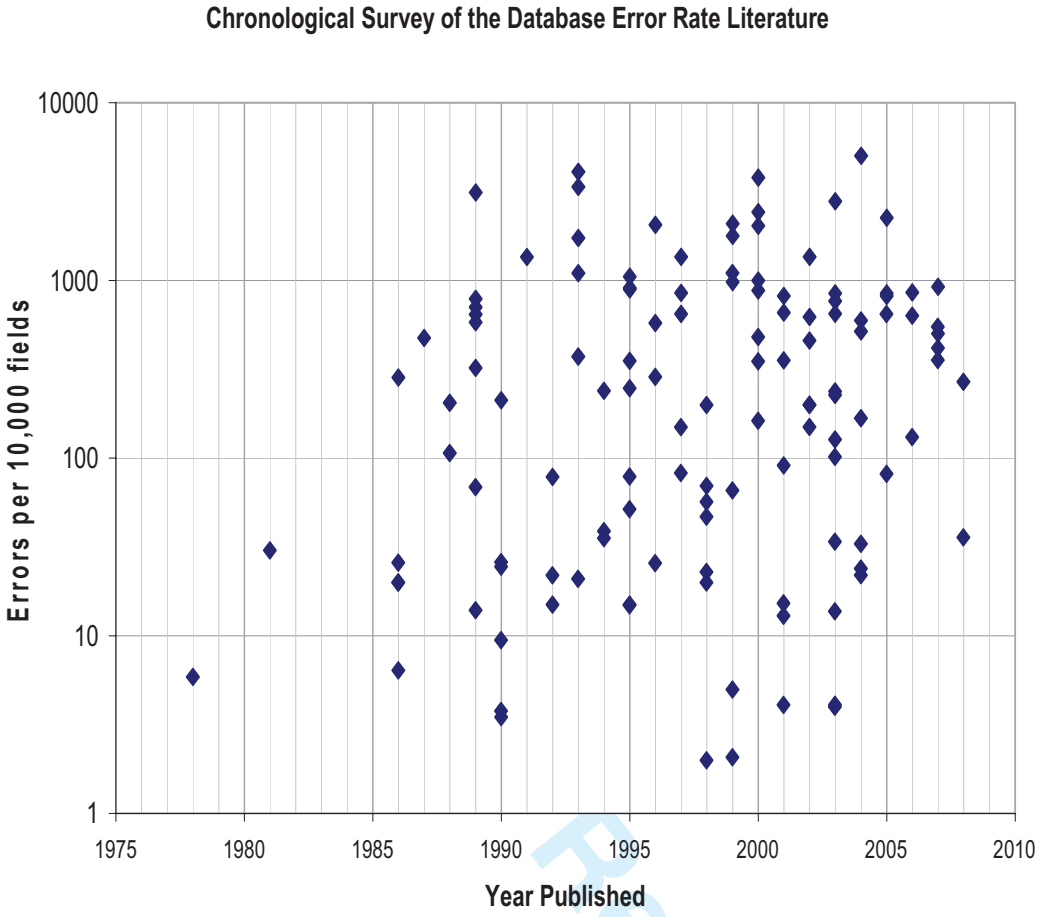
	Mean	Median	SD	Min	Max
Central	19	17	12	4	39
Distributed	28	24	17	15	69

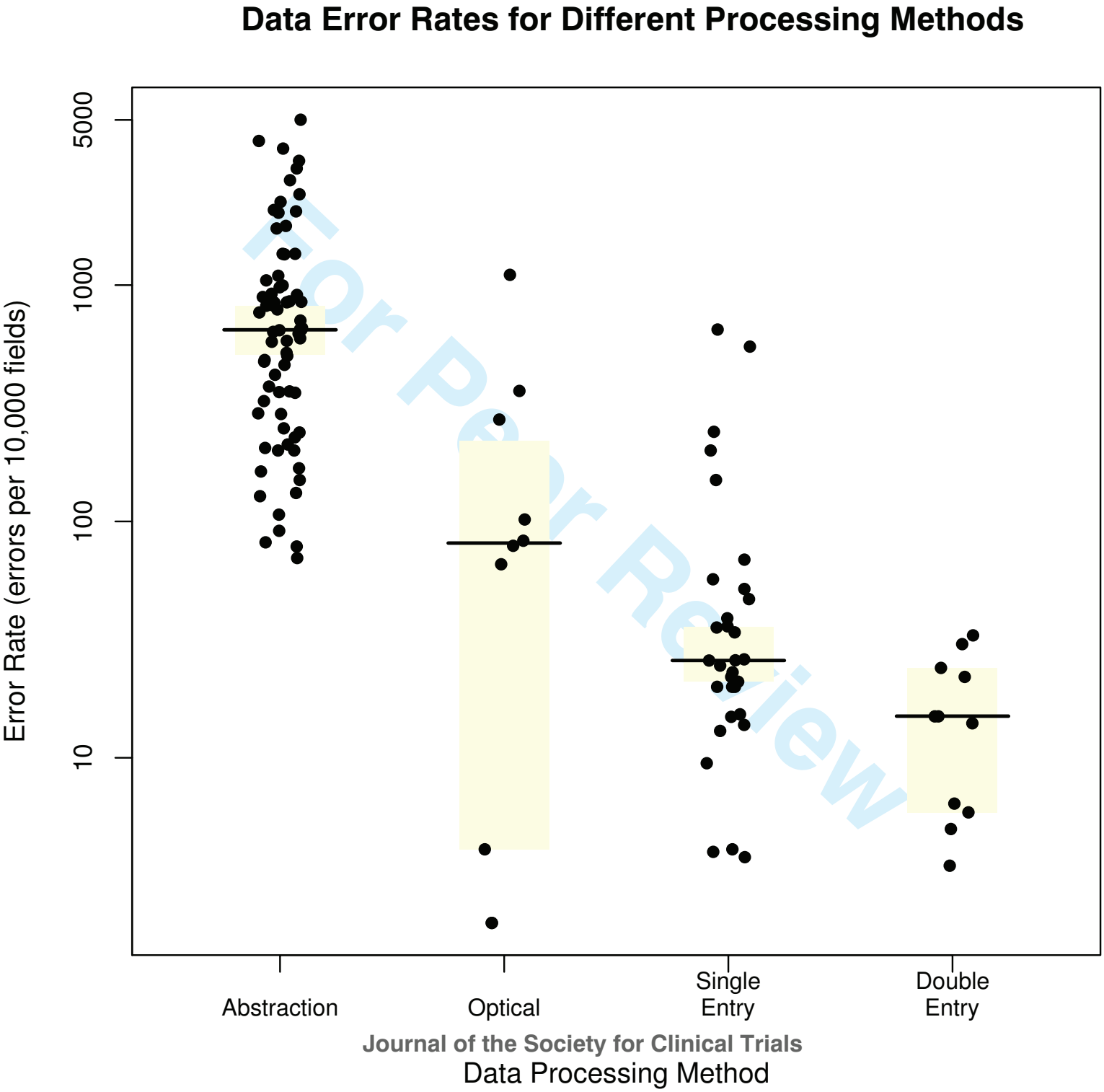
SD = standard deviation

Table 7 Descriptive statistics for batch subgroup analysis

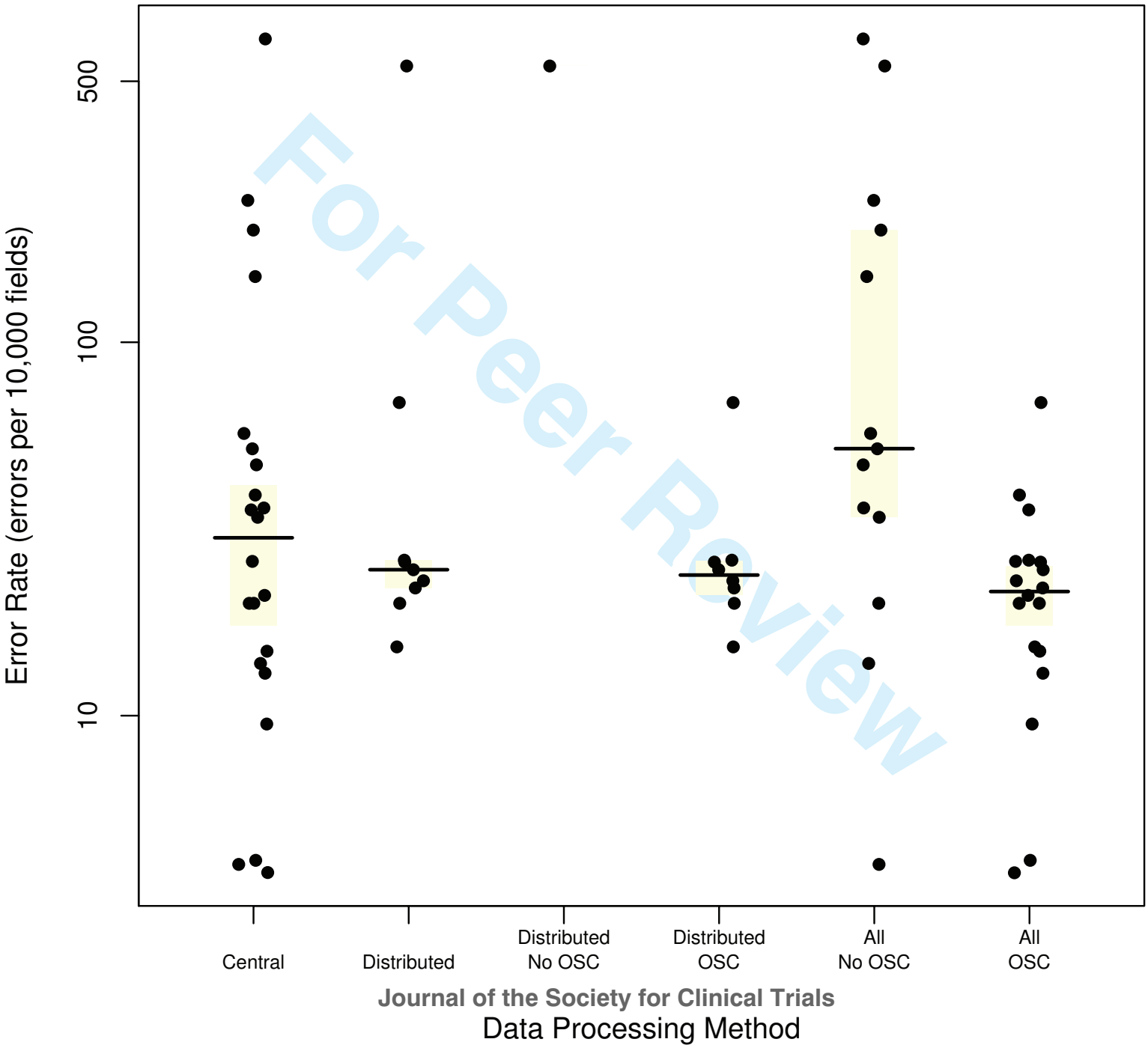
	Mean	Median	SD	Min	Max
No batch data cleaning reported	648	270	946	2	5019
Batch data cleaning reported	306	36	428	2	1351

SD = standard deviation

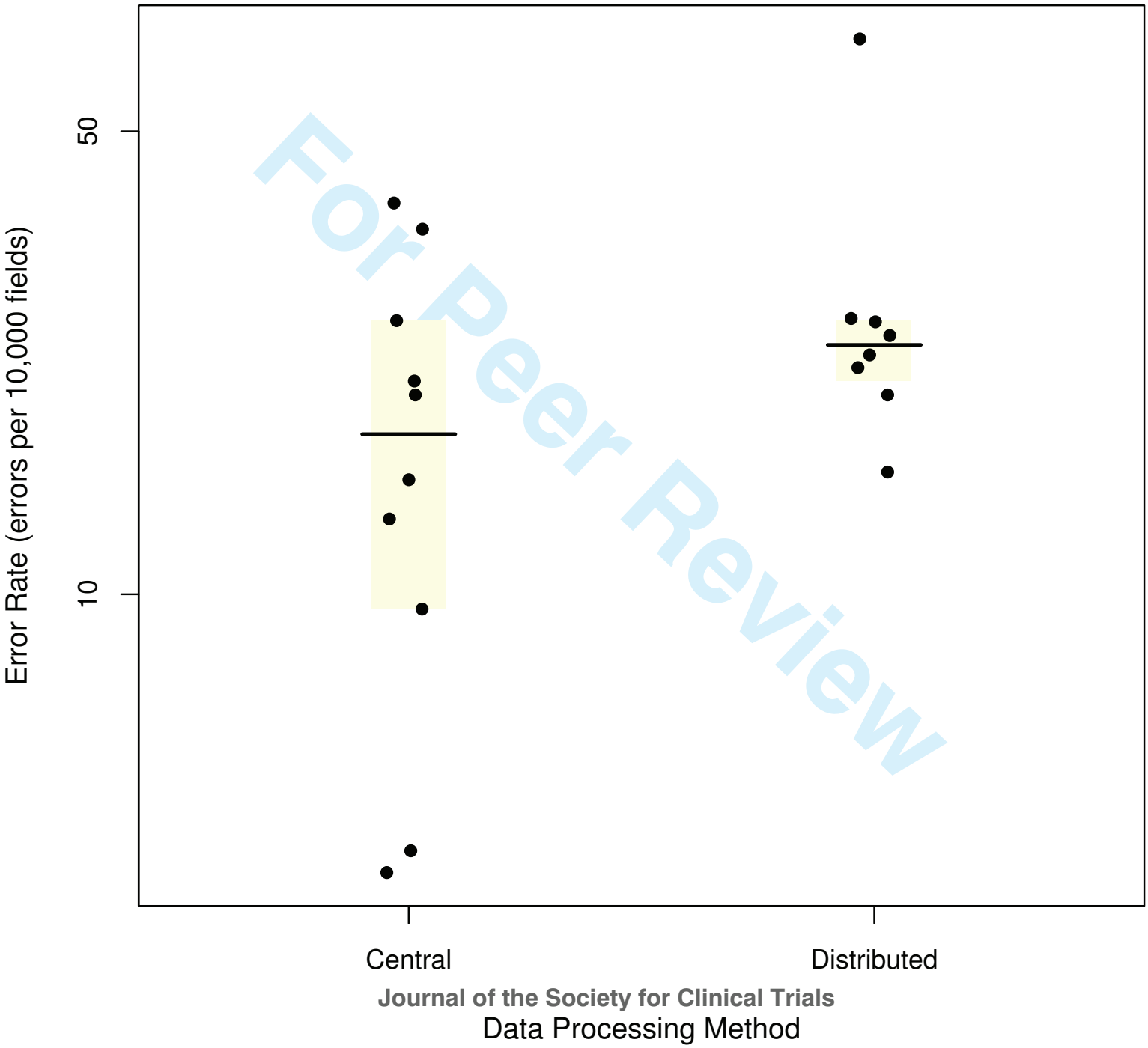

Chronological Survey of the Database Error Rate Literature
Errors per 10,000 fields
10000
1000
100
10
1
1975
1980
1985
1990
1995
2000
2005
2010
Year Published



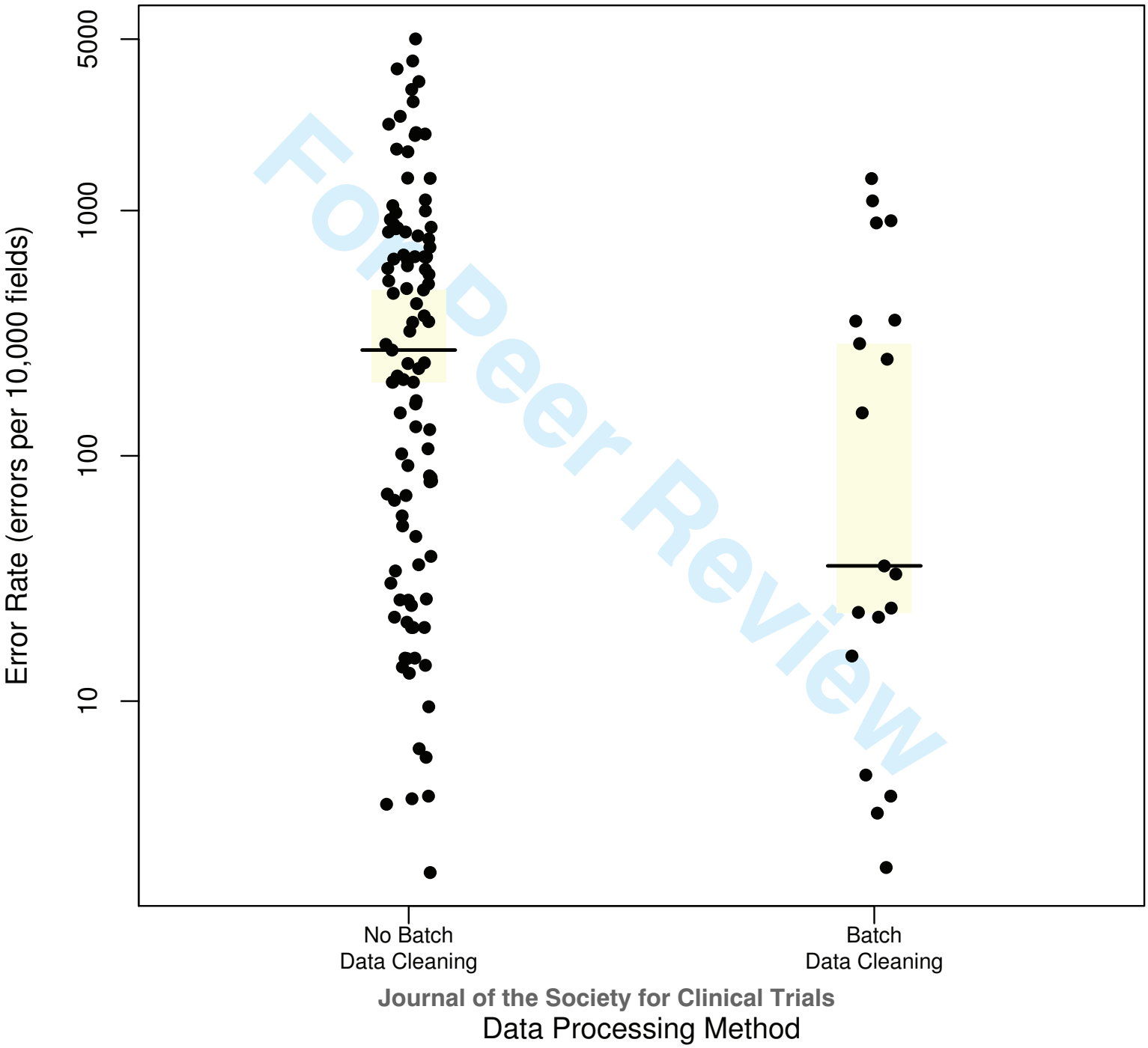
Subgroup Analysis for Single Entry



Error Rates from Central and Distributed Data Processing In Presence of On−Screen Checks



Subgroup Analysis for Batch Data Cleaning



Defining Data Quality for Clinical Research: A Concept Analysis

Journal:	***J*ournal of Data an*d* *I*nfor*m*ation Quality**
Manuscript ID:	draft
Manuscript Type:	Original Article
Keywords:Topics:	dIQ Impact, Strategy, policy < dIQ Impact
Keyword: Methodology:	Qualitative

Defining Data Quality for Clinical Research: A Concept Analysis

Despite notable previous attempts by experts to define data quality, the concept remains ambiguous and subject to the vagaries of natural language. This current lack of clarity continues to hamper research related to data quality issues. We present a formal concept analysis of data quality, which builds on and synthesizes previously published work. We further posit that discipline-level specificity may be required to achieve the desired definitional clarity. To this end, we combine work from the clinical research domain with findings from the general data quality literature to produce a discipline-specific definition and operationalization for data quality in clinical research. While the results are helpful to clinical research, the methodology of concept analysis may be useful in other fields to clarify data quality attributes and to achieve operational definitions.

Data Quality Concept Analysis

1. INTRODUCTION

The most common definition for quality, including data quality, is "fitness for use" [Batini and Scannapieco 2006; Lee, Pipino, Funk and Y. 2006]. This broad definition abstracts across differences among industries and focuses attention on the customer; both attributes are beneficial. However, the corresponding lack of definitional specificity precludes use in research, as well as operationalization within specific disciplines. Citing Juran's "fitness for use" definition, the American Society for Quality acknowledges both high-level and discipline-specific needs and defines quality as "a subjective term for which each person or sector has its own definition" [Quality Progress Editorial Staff 2008]. This formulation implies that employing a more abstract definition without refining its application within the specific discipline of data quality could have detrimental effects.

Successful theoretical development requires well-defined concepts. It is incumbent upon researchers within disciplines to define concepts with specificity capable of supporting scientific inquiry in the context of overarching theory. The purpose of this work is to clarify the term *clinical research data quality* for use in theory development, research, and application to operations. Simply put, we cannot test a poorly defined theory, nor manage what we cannot measure [Redman 1996].

2. BACKGROUND

2.1. Clouding Factors

There is ample reason for the fuzziness that has historically surrounded the concept of data quality. Most authors define data quality as a *multidimensional concept* with dimensions that include categories such as accuracy, currency, consistency, and completeness [Batini et al. 2004;

Batini and Scannapieco 2006; Pipino et al. 2002; Redman 1996; Tayi and Ballou 1998; Wand and Wang 1996; Wang and Strong 1996]. This multidimensionality (multiple attributes), however, can lead to ambiguity and imprecision because different groups may emphasize some dimensions while excluding others.

For example, the information technology (IT) sector defines data quality by conformity to data definition and stated business rules, while regulatory and legal groups tend to emphasize attribution and verifiability [U.S. Food and Drug Administration 2007]. It is likely that in many contexts only a subset of attributes will be considered important; thus, proliferation of use-related and use-specific definitional subsets will continue. Further complicating matters, data quality has dimensions that are *inherent to the data*, such as accuracy, and dimensions that are *context dependent*, such as timeliness and relevance [Lee et al. 2002; Wand and Wang 1996]. Moreover, even inherent characteristics may be *relative*; i.e., what constitutes good quality for some uses may be poor quality for others [Ge and Helfert 2007; Wand and Wang 1996].

Additional confusion exists between data quality and related concepts. For example, inherent characteristics of data (e.g., volatility and accessibility) have been used as dimensions of data quality. Likewise, attributes of the systems used to store and provide access to data are sometimes referred to as data quality dimensions, obscuring the distinction between data and system [Redman 1996]. For example, suboptimal system usability, implementation choices, or data structure may be construed as data quality issues.

2.2. Prior Attempts to Clarify the Concept of Data Quality

In spite of the efforts of experts in the field, the concept of data quality remains elusive. Some texts do not define data quality [Liepins and Uppuluri 1990; Maydanchik 2007; Naus 1975],

while other works adopt the generic (and abstract) "fitness for use" definition [Herzog et al. 2007; Olson 2003; Redman 1996; Redman 2001]. Lee and colleagues [2006] state that data quality should be defined in the context of the organization, using applicable dimensions from the literature, but they do not provide examples. Still others provide multiple definitions, or simply note that there "is no general agreement either on which set of dimensions defines data quality or on the exact meaning of each dimension" [Batini, Catarci and Scannapieco 2004].

Quality dimensions are similarly diverse and have been classified into three broad categories: 1) intuitive approach; 2) empirical approach; and 3) theory-based approach [Batini, Catarci and Scannapieco 2004], with each category ranked according to number of citations [Wand and Wang 1996]. The association of the same term to different definitions characterized by Batini [2006] indicates that progress toward a consensus dimension-based definition is still hindered by vagaries of natural language and idiosyncratic usage. Importantly, with the exception of Orr's System Theory work, Wand and Wang's ontological work, and Frank's ontological work specific to the geospatial mapping domain [Frank 2007; Orr 1998; Wand and Wang 1996], the concept of data quality has not been linked to theory.

Where specificity can not be obtained for an entire field (as in the case of data quality), it may be possible within a specific sector [Batini and Scannapieco 2006; Quality Progress Editorial Staff 2008]. Accordingly, we sought to achieve a discipline-specific definition and attribute list for clinical research data quality that accommodates the context of broader data quality work.

2.3 Needs of Clinical Research

Research data management is an exercise in mass customization [Koh 2008]. A clinical trial database is considered a product; the customers are the statisticians analyzing the data and the clinical community interpreting the results. The data collected and the level of quality appropriate for each variable both depend on the scientific question(s) being asked and the statistical tests applied in each study—one study's major independent and dependent variables and important covariates might be only background supporting data for another study. Necessary data quality also depends not only on which variable/s are in error, but also on the degree of robustness of the statistical analysis to those errors [Nahm et al. 2008].

Because the development of clinical knowledge often entails combining data from different studies, or from studies managed by different organizations, a consistent definition of data quality and standard ways of measuring it are critical [Nahm et al. 2004]. As these needs are discipline-specific, a correspondingly specific definition for clinical research data quality may be appropriate. Such a definition, however, should remain compatible with the context of the research done in the broader data quality domain.

3. METHODS

We performed a systematic concept analysis using the method of Walker and Avant, in which a concept is clarified through explication of antecedents, consequences, attributes, and cases [Walker and Avant 2005]. Due to the richness of work in data and information quality and our desire to preserve that broader context, we approached our project as a concept derivation rather than a *de novo* concept synthesis [Walker and Avant 2005]. Although *clinical research data*

quality shares dimensions with the more generic term used in IT and data warehousing, the importance and definition of the dimensions differ, as does the unit for decision-making.

Following Walker and Avant, attributes (called *dimensions* in the data quality literature), were identified through a review of the general data quality literature that used the Google Scholar™ search engine to identify relevant articles. Due to the large number of results retrieved (218,000 citations for *"data quality"* and 43,800 citations for *"data quality" AND "theory")*, we focused on review papers and published books. Attributes from the literature applicable to clinical research were selected and categorized. Where possible, literature definitions were used, although in some cases there were significant conflicts in definitions across various sources [Batini and Scannapieco 2006].

3.1. Uses of the Data Quality Concept

While business, IT, and manufacturing sectors tend to favor the "fitness for use" definition of data quality, the term has been used in clinical research to denote accuracy of the data; i.e., the extent to which the data reflect the true state of the patient. The Institute of Medicine defines quality data as those that "support the same conclusions as error-free data" [Institute of Medicine 1999]. In previous work, Nahm et al. applied the label data *quality* to what was in fact a measure of data *accuracy* [Nahm, Pieper and Cunningham 2008]. To a greater extreme throughout the clinical research industry, the result of a data quality assessment refers to a narrow scope, such as fidelity of data processing or compliance in measuring, observing, or collecting data [SCDM 2007]. In clinical registries, data quality refers to a subset of dimensions, most commonly accuracy and completeness [Arts et al. 2002]. In IT, data quality incorporates conformity to

syntactic constraints, as well as accuracy and completeness [Herzog, Scheuren and Winkler 2007; Naus 1975].

The term *quality* itself is problematic. Merriam-Webster's Dictionary lists eight definitions for quality (Figure 1), aspects of which are to be found in technical usage (e.g., degree of excellence; inherent feature; distinguishing attribute). Similarly, the definition of *data* remains a topic of academic debate [Johnson 2008]. One school of thought posits data as a difference (for example, a black dot on a white page constitutes data), and that data plus meaning create information [Floridi 2008]. Coiera, [2003] on the other hand, considers data the raw form of information.

[Figure 1. Caption set underneath illustration.]

As with the term *quality*, these differing definitions and interpretations cloud the term *data* (Figure 2).

[Figure 2. Caption set underneath illustration.]

For example, if we adopt the first Merriam-Webster's definition (factual information), the output of aggregate reports would constitute data, and thus the concept of data quality should apply at the aggregate level. Others, adopting a Floridian or Coieraian interpretation, likely would not apply the concept of data quality to an aggregate report. Additionally, *meaning* is a pivotal point between the Coieraian and Floridian definitions. One possible attribute of data quality is specificity (non-ambiguity), which requires meaning to accompany data, and implies that it is not data quality that we should be concerned with at all, but rather *information* quality.

Definitional work to date has not yielded the clarity needed for theoretical development and research. Because scientific and operational use of a concept requires focus and precision, this work proceeds with the scientific and operational uses of the term *data quality*.

3.2. Antecedents and Consequences

Data quality has a number of necessary precursors. First, there must be data, as defined by Floridi [2008] or Coiera [2003]. Second, data must possess attributes sufficient for us to recognize them as the particular data of interest. Interestingly, the more we know about data, the more they approach the Floridian definition of *information*; thus, the difference between data quality and information quality, if one exists, is not illuminated by this analysis. Third, there must be a representation of data such that humans or machines can recognize and process them. Fourth, there must be a source for comparison so that the quality or dimension of data quality can be judged (*i.e.,* a concept of truth, correct value, or reference point).

The ultimate consequence of data quality in the Juran sense is that the data are fit for the customer's use. In clinical research, data that possess quality "support the same conclusions as error [blemish] free data" [Institute of Medicine 1999]. (The term *blemish* was added to account for the dimensions in addition to accuracy that effect data use and usefulness.) Importantly, Juran's "fitness for use" definition is broader than the IOM's. In clinical trials, when a statistician decides whether available data can be used for an analysis, he or she is applying the IOM definition.

3.3. Dimensions

Data quality dimensions are data characteristics that affect use and usefulness; further, they are the mechanism through which we design, control, and increase data quality [Tayi and Ballout 1998]. The dimensions deemed necessary for clinical research—accuracy, currency, completeness, consistency, timeliness, relevance, granularity, specificity (non-ambiguity), for precision, and attribution were chosen from the literature (Table I) and used whenever possible.

[Table I. Table head on top of table.]

Where literature definitions did not provide an adequate fit with clinical research use, discipline-specific definitions were created. Also provided in Table I are operational definitions or metrics for definitions.

Of note, defining dimensions at the level of measurable metrics helped clarify a confounding factor characteristic of clinical research: while some dimensions may be assessed at the data value level (*e.g.*, correct vs. incorrect); in clinical research, decision-making occurs at the database level (*i.e.,* "Is this database of adequate quality to support the analysis?"). This database-level information is derived by aggregating value-level results (for example, number of errors divided by number of data values) and differs from the broader data quality methodology of value-level assessment and decisions.

As with the broader data quality field, clinical research requires both inherent and context-sensitive dimensions; for example, "timeliness" for expedited safety event data differs from the "timeliness" of database lock. We accommodated this issue by providing an operational definition or measure, while leaving the acceptance criterion context-specific. In the case of our example, "timeliness" is operationally defined as the difference between the date the data are needed and the date they are available. The acceptance criterion (i.e., the definition of "excessively late") is left to the customer. To clarify these different dimensions, Table I labels each dimension as inherent or context-sensitive and provides a context-sensitive operational definition.

Importantly, while inherent dimensions can be defined and assessed independently of context, operational definitions for context-sensitive dimensions depend upon the intended use. Thus, inherent dimensions can be stored as value-level or dataset-level metadata (*i.e.,* with data,

as advocated in several texts), whereas context-sensitive attributes by definition must be assessed at time of use [Batini and Scannapieco 2006; Lee, Pipino, Funk and Y. 2006].

This approach differs significantly from others in that the assessment, although counted on a data value-level basis, does not provide a data value-level metric for each dimension that can be stored with each data value. In clinical research, data quality is defined based on these ten dimensions as "dimensional metrics meeting the acceptance criteria of the data user and consumer."

3.4 Illustrating Cases

3.4.1 Ideal Case

A directory of research projects that contains descriptive information about each project was created as a searchable inventory to aid in identifying scientific collaborations. A researcher seeking potential collaborators for a cell therapy study queries the database and retrieves three projects. Two of the projects are hers, and she recognizes that the information is accurate, consistent, and complete. She also notes that her study coordinator has been maintaining the data and that it is current. The third project belongs to another researcher, indicating a potential collaborator. In this case, the data were available when needed, were relevant to her question, and provided a sufficient level of detail to answer her question; therefore, the data are of high quality.

3.4.2 Contrary Case

A researcher uses data from a similar directory at another institution to find potential collaborators for a cell therapy study. He queries the database and retrieves two projects, but

immediately suspects the data are incomplete, inaccurate, and inconsistent because one of his own projects is missing, and another is listed as belonging to the incorrect investigator. Upon closer examination, he is unable to tell who entered the erroneous data or when it was last updated. One of the two returned projects belonged to another researcher, pointing him toward a potential collaborator. However, it was unclear whether the telephone number listed was for the trial call center or the investigator. Not wanting to risk embarrassment, he decided not to attempt contacting the potential collaborator. In this case, although data were relevant and at an appropriate level of detail, they were inaccurate, unattributed, incomplete, and inconsistent, and therefore were not quality data.

3.4.3 Illegitimate Case

A researcher completes an analysis of a project and provides two tables and a listing to a colleague, who glances at the tables and replies, "Wow, nice job. These look great. High quality work!" This scenario is not a legitimate example of data quality, because no attributes were assessed. The colleague merely examined the table format and not the contents, and used no source of comparison for the assessment. In addition, aggregate numbers on a table are not always an indicator of data quality; they may be an indicator of programming quality.

3.4.4 Borderline Case

A researcher downloads transcripts from focus group meetings. After reading the information, he determines that the questions used in the groups are not sufficiently relevant to his research question and decides that he cannot use the data for his intended purpose. This is a borderline case, because relevance is an attribute of data quality. However, it is context dependent, in that

the same focus group data were relevant to the original research questions asked. Thus, data may not be relevant or at the right detail level for a secondary use, but may have been so for their initial use.

4.0 CONCLUSION

We have advanced a definition for data quality in clinical research that includes the following key attributes: accuracy, completeness, currency, consistency, timeliness, relevance, granularity, specificity, precision, and attribution meeting the acceptance criterion of the data user and consumer. Clarity has been added through explication, differentiation, and definition. This clarity was achieved at the expense of generality of application. An operationalizable definition for clinical research is a significant contribution, given its potential for supporting related research. We were able to use general concept analysis techniques to achieve the desired clarity regarding the concept of data quality in clinical research. This analysis, while subject to the same limitations of natural language that underlie the original ambiguities, delineates differentiating and orthogonal characteristics of data quality attributes (e.g., inherent versus context-sensitive, that enable operational definitions. Thus, while the results are helpful to clinical research, the methodology of concept analysis may be useful in other fields to clarify data quality attributes and to achieve operational definitions.

REFERENCES

ARTS, D.G., DEKEIZER, N.F. AND SCHEFFER, G.J. 2002. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association 9*, 12.

BATINI, C., CATARCI, T. AND SCANNAPIECO, M. 2004. A Survey of Data Quality Issues in Cooperative Information Systems. In *Proceedings of the 23rd International Conference on Conceptual Modeling, ER.*, Shanghai.

BATINI, C. AND SCANNAPIECO, M. 2006. *Data Quality Concepts, Methodologies and Techniques*. Springer-Verlag, Berlin.

COIERA, E. 2003. *Guide to Health Informatics* Arnold Publishers, Australia.

FLORIDI, L. 2008. Semantic Conceptions of Information. In *The Stanford Encyclopedia of Philosophy*, E.N. ZALTA Ed. Stanford University.

FRANK, A.U. 2007. Incompleteness, Error, Approximation, and Uncertainty: An Ontological Approach to Data Quality. In *Geographic Uncertainty in Environmental Security*, M. A. AND S. KOKHAN Eds. Springer.

GE, M. AND HELFERT, M. 2007. A Review of Information Quality Research - Developing a Research Agenda. In *Proceedings of the International Conference on Information Quality*, Cambridge, Massachusetts, November 9 2007.

HERZOG, T.N., SCHEUREN, F.J. AND WINKLER, W.E. 2007. *Data Quality and Record Linkage Techniques*. Springer, New York.

INSTITUTE OF MEDICINE 1999. *Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making*. National Academy Press, Washington DC.

JOHNSON, T.R. 2008. What are Data, Information, and Knowledge and What is Their Role in Biomedical Informatics? University of Texas at Houston School of Health Information Sciences.

KOH, K. Personal communication, email from Karen Koh to Meredith Nahm dated December 3, 2008.

LEE, Y.W., PIPINO, L.L., FUNK, J.D. AND Y., W.R. 2006. *Journey to Data Quality*. MIT Press, Cambridge, MA.

LEE, Y.W., STRONG, D.M., KAHN, B.K. AND WANG, R.W. 2002. AIMQ: A Methodology for Information Quality Assessment. *Information & Management 40*, 14.

LIEPINS, G.E. AND UPPULURI, V.R.R. 1990. Data Quality Control Theory and Pragmatics Marcel Dekker, New York.

MAYDANCHIK, A. 2007. *Data Quality Assessment*. Technics Publications, New Jersey.

NAHM, M., DZIEM, G., FENDT, K., FREEMAN, L., MASI, J. AND PONCE, Z. 2004. Data Quality Survey Results. *Data Basics 10*, 7.

NAHM, M.L., PIEPER, C.F. AND CUNNINGHAM, M.M. 2008. Quantifying data quality for clinical trials using electronic data capture. *PLoS ONE 3*, e3049.

NAUS, J. 1975. *Data Quality Control and Editing*. Marcel Dekker, New York.

OLSON, J.E. 2003. *Data Quality the Accuracy Dimension*. Morgan-Kaufmann Publishers, San Francisco.

ORR, K. 1998. Data Quality and Systems Theory. *Communications of the ACM 41*, 6.

PIPINO, L., LEE, Y. AND WANG, R. 2002. Data Quality Assessment. *Communications of the ACM 45*, 8.

QUALITY PROGRESS EDITORIAL STAFF 2008. American Society for Quality Glossary. In *American Society for Quality Glossary* American Society for Quality, Milwaukee, WI

REDMAN, T.C. 1996. *Data Quality for the Information Age*. Artech House, Boston, MA.

REDMAN, T.C. 2001. *Data Quality Field Guide*. Digital Press, Boston, MA.

SCDM 2007. Good Clinical Data Management Practices Society for Clinical Data Management, www.scdm.org.

TAYI, G.K. AND BALLOU, D.P. 1998. Examining Data Quality. *Communications of the ACM 41*, 4.

U.S. FOOD AND DRUG ADMINISTRATION, U.S.D.O.H.A.H.S. 2007. Guidance for Industry, Computerized Systems Used in Clinical Trials, D.O.H.A.H. SERVICES Ed. U.S. Food and Drug Administration, Rockville, MD.

WALKER, L.O. AND AVANT, K.C. 2005. *Strategies for Theory Construction in Nursing*. Pearson Education, Inc., Upper Saddle River, New Jersey.

WAND, Y. AND WANG, R. 1996. Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM 39*, 10.

Data Quality Concept Analysis

WANG, R. AND STRONG, D. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems 12*, 30.

Data Quality Concept Analysis

Figure Legends

Figure 1. Definitions of Quality. From: Merriam-Webster Online Dictionary, 2008 (used with permission).

Figure 2. Definitions of Data. From: Merriam-Webster's Online Dictionary, 2008 (used with permission).

Data Quality Concept Analysis

Table 1. Data Quality Dimensions for Clinical Research

	Inherent	Context sensitive	Natural Language Definition	Operational Definition / Metric
Accuracy	Yes	No	*States in the data match the intended state in the real world* (data values represent real world truthfully)	Number of errors divided by number of fields inspected (implies comparison with a gold standard)
Currency	Yes	No	The length of storage for a data value (since last update)	Date of use/need minus date data last updated
Completeness	Yes	No	*The extent to which every represented real world state is reflected in the data*	Number of missing values divided by number of fields assessed
Consistency (internal)	Yes	No	*Where there are multiple representations of real world states in the data, they are the same.* For clinical research, extended to: data values representing the same real-world state are not in conflict	Number of discrepant values divided by number of values subject to data consistency checks
Timeliness	No	Yes	*Length of time from a change in the real world state to the time when the data reflect the change*	Date data are needed minus date data are ready for intended use
Relevance	No	Yes	Data can be used to answer a particular question	Percentage of data values applicable to intended use
Granularity	No	Yes	Level of detail captured in data	Percentage of values at level of detail appropriate for intended use
Specificity (non-ambiguity)	Yes	No	*Each state in the data definition* (metadata) *corresponds to one* (or none) *state of the real world*	Number of values with full ISO 11179 metadata, including definition, divided by number assessed
Precision	No	Yes	Number of significant digits to which a continuous value was measured (and recorded); for categorical variables, the resolution of the categories	Percentage of values with precision appropriate for intended use
Attributability	Yes	No	Source and individual who generates and updates data are inextricably linked to data values	Percentage of data values linked to source and user ID of person who generated and changed the record

Italicized wording quoted from Wand and Wang 1996.

Figure 1. Definitions of Quality. From: Merriam-Webster Online Dictionary, 2008 (used with permission).

Figure 1: Definitions of Quality

1 a: peculiar and essential character **:** NATURE <her ethereal *quality*> **b**: an inherent feature **:** PROPERTY <had a *quality* of stridence, dissonance> **c: CAPACITY** , ROLE <in the *quality* of reader and companion>

2 a: degree of excellence **:** GRADE <the *quality* of competing air service> **b**: superiority in kind <merchandise of *quality*>

3 a: social status **:** RANK **b**: ARISTOCRACY

4 a: a distinguishing attribute **:** CHARACTERISTIC <possesses many fine *qualities*> b*archaic* **b** : an acquired skill **:** ACCOMPLISHMENT

5: the character in a logical proposition of being affirmative or negative

6: vividness of hue

7 a: TIMBRE **b**: the identifying character of a vowel sound determined chiefly by the resonance of the vocal chambers in uttering it

8: the attribute of an elementary sensation that makes it fundamentally unlike any other sensation

--Merriam-Webster Online Dictionary

Figure 2. Definitions of Data. From: Merriam-Webster's Online Dictionary, 2008 (used with permission).

Figure 2: Definitions of Data

1 : factual information (as measurements or statistics) used as a basis for reasoning, discussion, or calculation <the data is plentiful and easily available> <comprehensive data on economic growth have been published>

2 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

3 : information in numerical form that can be digitally transmitted or processed

--Merriam-Webster Online Dictionary

Medical Record Abstractor's Perceptions of Factors Impacting the Accuracy of Abstracted Data

Meredith Nahm, MS[1, 2], Constance M. Johnson, PhD[3], Todd Johnson, PhD[1], Amy Franklin, PhD[1], Jack Smith, PhD[1] Jiajie Zhang, PhD[1]

[1]University of Texas, School of Health Information Sciences, Houston TX, [2]Duke Translational Medicine Institute, Durham, NC, [3]Duke University School of Nursing

Abstract

Medical record abstraction (MRA) is known to be a significant source of data errors in secondary data uses. Factors impacting the accuracy of abstracted data are not reported consistently in the literature. Two Delphi processes were conducted with experienced medical record abstractors to assess abstractor's perceptions about the factors. The Delphi process identified 9 factors that were not found in the literature, and differed with the literature by 5 factors in the top 25%. The Delphi results refuted seven factors reported in the literature as impacting the quality of abstracted data. The results provide insight into and indicate content validity of a significant number of the factors reported in the literature. Further, the results indicate general consistency between the perceptions of clinical research medical record abstractors and registry and quality improvement abstractors.

Introduction

As early as 1969, medical record abstraction was associated with poorly described processes, inconsistency and error[1]. By 1990, although many advancements in data collection for research had occurred, MRA was still regarded as problematic[2]. While MRA remains a common method of data collection for research and secondary data use[3-6], recent publications highlight the persistence of data accuracy problems[7, 8].

Importantly, MRA has been associated with the highest and most variable error rates of common data collection and processing methods[8]. These error rates are high enough to cause problems in the use of abstracted data[8-11].

This study was conducted to characterize abstractor perceptions about what factors impact the accuracy of abstracted data. Research questions included: 1) What factors do experienced abstractors think impact the accuracy of abstracted data? 2) Does experience as a clinical research or registry / quality improvement abstractor create differences in perception? 3) Do abstractor perceptions differ from those factors most frequently mentioned in the literature? If so, what are the differences?

Background

The literature contains several hundred articles about medical record abstraction, however, existing work is largely a-theoretical. Further, an authoritative definition of MRA has not been articulated. The following operational definition of MRA is used for this research: A process in which a human manually searches through a medical record to identify data required for a secondary use. Abstraction involves some direct matching of information found in the record to the data required, but also includes operations on the data such as categorizing, coding, transforming, interpreting, summarizing, and calculating. The abstraction process results in a summary of information about a patient for a specific secondary data use.

While at first blush, extraction of data directly from electronic medical records may appear to be the solution to accuracy issues in MRA, there are many reasons why this is not likely the case[12]. Computer programming resources, and sophisticated data query tools, are currently required for such extraction. These resources are not available to many researchers, and few, if any, tools exist to automate components of the abstraction process. Further, barriers to accessing such resources may be high, and for small investigator initiated studies, the costs likely exceed those of manual abstraction. In these cases, MRA will still be used and data accuracy from MRA will remain a concern.

Methods

Two concurrent Delphi[13] processes were conducted, one with experienced clinical research abstractors and one with experienced registry and quality improvement abstractors to obtain from the expert abstractors a list of factors that impact accuracy of abstracted data.

Abstractors for the clinical research Delphi were recruited at the Society for Clinical Research

Associates (SoCRA) national conference in September 2009. Registry and quality improvement abstractors were recruited at the American Health Information Management Association (AHIMA) National Convention in September 2009.

Eligible participants were individuals having:

1. Three or more years of abstraction experience as reported by the participant
2. Abstraction experience in either a clinical research or registry / quality improvement setting
3. Able and willing to give informed consent

In the first round, participants were asked to list from five to ten factors that based on their experience, impacted the accuracy of abstracted data. Following Round 1, the factors were reviewed and sorted according to semantic matches to obtain a list of distinct factors and the number of times each was mentioned.

The Round 1 factors were combined with factors obtained from a systematic literature review and provided back to the participants in Round 2. Factors were presented as statements that each increased or decreased the accuracy of abstracted data. In Round 2, the participants were asked to rate their level of agreement with these statements on a five point likert scale (strongly disagree, mildly disagree, neither agree or disagree, mildly agree, and strongly agree). Only participants who completed Round 2 were eligible to participate in subsequent rounds.

In Round 3, the participants in each Delphi were each provided an individualized report of their Round 2 responses versus the aggregate responses of their Delphi. In Round 3, participants were 1) asked for more information about factors where their response was within one point[*] of the aggregate and factors where their responses differed by more than one point from the aggregate, and 2) given the opportunity to change their responses should they wish to do so. Requesting participants to tell the interviewer more about their responses enabled researchers to assure consistent understanding of the statements on the Round 2 questionnaire as well as provided the researchers more in depth information about factors where there were disagreements, *i.e.,* where answers depended on things external to the statement such as differences in study types, or clinical area.

[*] One point was chosen because a difference of one point is the difference between the categories on the likert scale.

In Round 4 the participants were each provided an individualized report of their Round 3 responses and the aggregate of their Delphi. In all other aspects, Round 4 was conducted in the same manner as Round 3.

Rounds 1 and 2 were conducted using the cogix web-based survey system. Rounds 3 and 4 were conducted via structured phone interview. To prevent bias, an interviewer independent from the study team was used.

Twenty clinical research abstractors and 18 registry and quality improvement abstractors were consented to participate in this research to retain at least seven participants in the last round[13]. The participation rates for the rounds are shown in Table 1.

	Delphi Round			
	1	**2**	**3***	**4***
Clinical Research	80%	85%	68%	68%
Registry / QI	83%	67%	88%	88%

* denominator for Rounds 3 and 4 is the number of participants in Round 2.

Table 1. Delphi Participation Rates

This research was reviewed and approved by the Duke University and University of Texas Institutional Review Boards.

Systematic Literature Review

To inform Round 2, a systematic review of the literature was conducted to identify factors as impacting the accuracy of abstracted data. A PubMed query[†] in October of 2009 identified 361 articles. Abstracts for the 361 articles were reviewed by two people resulting in the exclusion of 287. The 74 included articles were read in full by two independent people. One hundred and thirteen new articles and technical reports were identified from citations in the reviewed articles, resulting in 187 articles reviewed. Of the 187 reviewed articles, 37 were subsequently found not to meet the inclusion criterion, leaving 150 articles ultimately included in the review. The inclusion criterion for this review were that the articles had to be in the English

[†] (((abstraction[Title/Abstract]) OR ("chart review"[Title/Abstract])) OR ("medical record review"[Title/Abstract])) AND ("clinical trial"[Title/Abstract] OR registry[Title/Abstract] OR "clinical research"[Title/Abstract] OR quality[Title/Abstract] OR "performance"[Title/Abstract]) AND (error[Title/Abstract] OR accuracy[Title/Abstract] OR "data quality"[Title/Abstract] OR errors[Title/Abstract] OR decision[Title/Abstract] OR reliability[Title/Abstract] OR validity[Title/Abstract])

language, describe use of healthcare data with the medical record as the source, and mention at least one factor in the accuracy of medical record abstraction.

Each reviewer independently read each included article and identified all statements of things that impact (increase, decrease, or stated without valence) accuracy of data abstracted from medical records. Each reviewer was instructed to combine semantically similar factors as such, *e.g.,* reabstraction and independent re-review of charts were counted as one factor rather than two distinct factors. Factors stated at different levels of granularity or with different modifiers or context, *e.g.*, reabstraction versus *ongoing* reabstraction were retained as separate factors. Factors stated with opposing valence were retained as distinct factors, *e.g.,* training abstractors increases accuracy versus lack of training decreases accuracy. Independent factor lists generated by each reader were compared and disagreements were resolved through discussion. This initial review identified 309 unique factors from 1063 instances of mentioned factors in the literature. The list of distinct factors was used to categorize the Delphi Round 1 results.

The factors were sorted by frequency of mention. Seventy five (24%) of the literature factors had greater than three mentions. These factors were combined with the factors obtained from Round 1 of the Delphi and were provided together to the participants in Round 2. Factors from the literature were injected into the Delphi process at Round 2 rather than Round 1 to prevent bias, *e.g.,* participants agreeing with stated factors rather than providing factors most important from their experience. In addition to standardized definitions, literature factors were introduced to reduce variability that could arise from different individual mental models of abstraction, *e.g.,* some individuals consider case ascertainment as part of abstraction while others do not. The number of factors from the literature used in Round 2 was limited by the time allotted for the rounds, because the informed consent stated that each Round would take less than an hour.

Member checking occurred as part of the Delphi design, *i.e.,* participants see the aggregate results of the previous Delphi round. A peer debriefing session was conducted with 30 independent study coordinators from Duke University Medical Center in February 2010.

Round 1 Results

Round 1 of the Delphi returned 227 instances of mentioned factors from which 94 distinct factors receiving one or more mention were identified. Twenty seven (29%) of the distinct factors received greater than two mentions. Six mentioned items fell outside of the working definition of Medical Record Abstraction. Five mentioned items could not be classified or otherwise labeled due to ambiguity of the information provided by the participant. Nine factors identified in Round 1 were not mentioned in the literature at all (Table 2). Five factors were mentioned in the literature, but were not in the top 24% (Table 3).

Factor	Number of Mentions
Abstractor credentials	10
Access to charts	6
Interruptions	6
Complete and accurate medical record	4
Availability of abstraction tools	4
Adequate time for abstraction tasks	4
Complexity of the study or project	3
Supportive collegial relationships with physicians, nurses, and medical records colleagues	3
Abstractor (human) error	3

Table 2. Factors identified in Delphi Round 1 that were Not Found in the Literature

Factor	Number of Mentions Delphi / Literature
Limited time	5 / 1
Lack of training	4 / 2
Same information found in multiple places in the medical record (opportunity for conflicting information)	3 / 3
Incomplete review of the medical record	3 / 1
Volume of information in the medical record	3 / 1

Table 3. Factors identified in Delphi Round 1 Not in the Literature top 24 %

A total of 89 distinct factors (the top 29% of the Round 1 responses and the top 24% of the literature factors) were carried forward to Round 2. The top 29% and 24% correspond to > 2 and > 3 mentions respectively. Including all mentions in each category, *i.e.,* all factors with >2 (Round 1) or >3 (literature) mentions, caused the difference in percentages. Three semantically redundant items were added to test internal consistency, seven items were created from existing factors containing multiple concepts; 99 items were used in Round 2 of the Delphis.

Round 1 Discussion

Five of the factors not mentioned in the literature were combinable with higher level factors or were related to factors mentioned in the literature. For example, abstractor credentials received 10 mentions in the Delphi, while the literature contained mentions for "necessity of a RN", "presence of an advanced degree", and "certification of abstractors"[14-17]. For this analysis factors at different levels of granularity were not combined.

"Adequate time for abstraction tasks" mentioned in the Delphi round 1, and "limited time" mentioned in both the Delphi round 1 and the literature were not combined. Likewise, "limited time" was not combined with "lack of resources". "Access to charts" (Delphi Round 1), although related to "missing charts" (Literature) was not combined due to incomplete concept overlap and opposite valence. Similarly, "complete and accurate medical record" (Delphi Round 1) was not combined with "existing error in the medical record" (Literature). "Availability of abstraction tools" (Delphi Round 1) was not combined with any of the numerous more granular literature factors that mentioned different types of abstraction tools, *e.g.,* data collection form, coding conventions, or data element definitions. The afore mentioned five factors, while exhibiting differences with factors reported in the literature are not considered by the authors as new factors identified through the Delphi.

While the literature mentioned several instances of errors that the authors would classify as human error, *e.g.,* "abstractor overlooked values in the medical record", these instances were not classified to the universal human error by the authors. Human error and other factors considered by the authors as newly identified by the Delphi are bolded in Table 2.

Delphi Results (all rounds)

Combining both Delphis, there were 73 factors with overall average ratings between mildly and strongly agree, 75 registry & quality improvement, 71 clinical research. There were 3 factors with an overall rating lower than neutral, all of these were between mildly disagree and neutral. The registry and quality improvement Delphi rated 7 factors lower than neutral (Table 4). The clinical research Delphi rated 2 factors lower than neutral. All factors rated lower than neutral originated from the literature.

Factor	CR	R / QI	Overall
RN credential	3.2	2.2	2.8
Blinding abstractors to study aims	2.5	1.9	2.2
Centralized abstraction	3.2	2.7	3.0
High study / project complexity	3.8	2.5	3.3
Thick medical records	3.2	2.8	3.1
Patients cared for by multiple providers	3.6	2.8	3.4
Presence of multiple diagnoses / procedures	2.8	2.6	2.7

Table 4. Factors Rated Lower Than Neutral

Discussion of Delphi Results (all rounds)

The two strongly refuted factors, "necessity of the Registered Nurse credential", and "blinding of abstractors" were contentious in the literature. Some argued for necessity of the RN credential due to the associated knowledge of data flow and documentation in the healthcare environment, ability to locate information in the medical record, and fluency in medical language[6, 14-17]. Others argued that individuals with clinical knowledge were more apt to interpret information in the medical record rather than rigidly follow abstraction guidelines. Having the RN credential strongly correlated with the perceived necessity of the credential for MRA.

The literature was similarly conflicted regarding blinding of abstractors. Some argued that blinding abstractors to study endpoints prevented bias[16, 18, 19]. Others argued that knowledge of the study purpose and endpoints was necessary for abstractors to do a good job[7, 17, 20]. This difference in opinion may be due to different perceptions between investigators versus abstractors, or different application areas, *e.g.,* explicit versus implicit abstraction.

The low rating for presence of multiple diagnoses or procedures is puzzling. This factor is cited by over 8

articles in the clinical research, registry and quality improvement literature as a factor impacting the accuracy of abstracted data. The reported mechanism is the difficulty assigning a primary diagnosis from multiple possible problems. Further, two large and robust studies in abstraction for billing conducted in 1977 by the IOM reported this as a major finding[21, 22]. It is possible that this is no longer a factor, or that while it may be a significant factor in claims abstracting, it is not problematic in other areas.

Limitations

Homogeneity of participants is a critical factor in the Delphi process. Our Delphi participants were homogenous with respect to abstraction setting (clinical research versus registry and quality improvement) and experience level. However, there is significant variation of practice within each setting. Further, factors that may impact medical record abstraction may vary across clinical area. Thus, these results should be assessed in a particular practice area rather than blindly applied. Additionally, over 200 factors mentioned in the literature, *i.e.,* the bottom 76%, could not be evaluated in this research.

Conclusion

From the consistency between the two Delphis, we conclude that the factors impacting accuracy are similar, *i.e*., differences between abstraction setting is not itself among the largest of factors, and that best practices and methods that improve accuracy in abstraction for registries and quality improvement projects may be applicable to clinical research.

From the number of factors and the high level (73%) of agreement between expert abstractors and the literature, we conclude that data accuracy in MRA is a complex, many-faceted problem. Thus, solutions to improving, controlling and assuring accuracy of abstracted data will necessarily be multi-faceted.

The project received support from grant numbers UL1RR024128 and UL1RR024148 to Duke Univ. and Univ. of Texas Health Science Center Houston.

References

1. Feinstein AR, Pritchett JA, Schimpff CR. The epidemiology of cancer therapy. IV. The extraction of data from medical records. *Arch Intern Med.* May 1969;123(5):571-590.
2. Collen MF. Clinical research databases--a historical review. *J Med Syst.* Dec 1990;14(6):323-344.
3. Pan L, Fergusson D, Schweitzer I, Hebert PC. Ensuring high accuracy of data abstracted from patient charts: the use of a standardized medical record as a training tool. *J Clin Epidemiol.* Sep 2005;58(9):918-923.
4. Thoburn KK, German RR, Lewis M, Nichols PJ, Ahmed F, Jackson-Thompson J. Case completeness and data accuracy in the Centers for Disease Control and Prevention's National Program of Cancer Registries. *Cancer.* Apr 15 2007;109(8):1607-1616.
5. Worster A, Bledsoe RD, Cleve P, Fernandes CM, Upadhye S, Eva K. Reassessing the methods of medical record review studies in emergency medicine research. *Ann Emerg Med.* Apr 2005;45(4):448-451.
6. vonKoss Krowchuk H, Moore ML, Richardson L. Using health care records as sources of data for research. *J Nurs Meas.* Summer 1995;3(1):3-12.
7. Engel L, Henderson C, Fergenbaum J, Colantonio A. Medical record review conduction model for improving interrater reliability of abstracting medical-related information. *Eval Health Prof.* Sep 2009;32(3):281-298.
8. Nahm M, Johnson CM, Johnson TR, Fendt K, Zhang J. Clinical Reserarch Data Quality Literature Review and Systematic Analysis. *Clinical Trials Journal.* submitted.
9. Freedman LS, Schatzkin A, Wax Y. The impact of dietary measurement error on planning sample size required in a cohort study. *Am J Epidemiol.* Dec 1990;132(6):1185-1195.
10. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biol Psychiatry.* Apr 15 2000;47(8):762-766.
11. Rostami R, Nahm M, Pieper CF. What can we learn from a decade of database audits? The Duke Clinical Research Institute experience, 1997--2006. *Clin Trials.* Apr 2009;6(2):141-150.
12. Lowenstein SR. Medical record reviews in emergency medicine: the blessing and the curse. *Ann Emerg Med.* Apr 2005;45(4):452-455.
13. Linstone HA, Turoff M, eds. *The Delphi Method Techniques and Applications.* web ed2002.
14. Eder C, Fullerton J, Benroth R, Lindsay SP. Pragmatic strategies that enhance the reliability of data abstracted from medical records. *Appl Nurs Res.* Feb 2005;18(1):50-54.
15. Hemmila MR, Jakubus JL, Wahl WL, et al. Detecting the blind spot: complications in the trauma registry and trauma quality improvement. *Surgery.* Oct 2007;142(4):439-448; discussion 448-439.
16. Cassidy LD, Marsh GM, Holleran MK, Ruhl LS. Methodology to improve data quality from chart review in the managed care setting. *Am J Manag Care.* Sep 2002;8(9):787-793.
17. Simmons B, Bennett F, Nelson A, Luther SL. Data abstraction: designing the tools, recruiting and training the data abstractors. *SCI Nurs.* Spring 2002;19(1):22-24.
18. Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: Where are the methods? *Ann Emerg Med.* Mar 1996;27(3):305-308.
19. Badcock D, Kelly AM, Kerr D, Reade T. The quality of medical record review studies in the international emergency medicine literature. *Ann Emerg Med.* Apr 2005;45(4):444-447.
20. Reisch LM, Fosse JS, Beverly K, et al. Training, quality assurance, and assessment of medical record abstraction in a multisite study. *Am J Epidemiol.* Mar 15 2003;157(6):546-551.
21. Demlo LK. Reliability of hospital; discharge abstracts. In: Medicine Io, edWashington DC: National Academy of Sciences; 1977.

22. Demlo LK. Reliability of medicare hospital discharge records: report of a study In: Medicine Io, edWashington DC: National Academy or Sciences; 1977.

Distributed Cognition Artifacts on Clinical Research Data Collection Forms

Meredith Nahm, MS[1, 2], Vickie D. Nguyen, MA[1], Elie Razzouk, MD[1], Min Zhu, MD, MS[1], Jiajie Zhang, PhD[1]
[1]University of Texas, School of Health Information Sciences, Houston, TX,
[2]Duke Translational Medicine Institute, Durham, NC

Abstract

Medical record abstraction, a primary mode of data collection in secondary data use, is associated with high error rates. Distributed cognition in medical record abstraction has not been studied as a possible explanation for abstraction errors. We employed the theory of distributed representation and representational analysis to systematically evaluate cognitive demands in medical record abstraction and the extent of external cognitive support employed in a sample of clinical research data collection forms.

We show that the cognitive load required for abstraction in 61% of the sampled data elements was high, exceedingly so in 9%. Further, the data collection forms did not support external cognition for the most complex data elements. High working memory demands are a possible explanation for the association of data errors with data elements requiring abstractor interpretation, comparison, mapping or calculation. The representational analysis used here can be used to identify data elements with high cognitive demands.

Introduction

Data collection in clinical research, both retrospective and prospective, relies on the abstraction of data from medical records[1, 2]. Abstraction is a time and resource intensive task[3, 4] and is associated with high error rates[5]. However, little is known about the causes and mitigators of these errors[6]. Over time, authors have suggested that the design of the data collection form is a significant factor in the accuracy of abstracted data[7, 8, 9, 10]. Although data collection forms are widely touted as a key factor in data quality, little evaluative work has been done to understand the mechanism and impact of data collection form design on data accuracy. Today, the design of data collection forms is guided by primarily a-theoretical lists of things that form designers should and should not do[8, 15, 16].

While the role of paper-based patient records in clinician cognition has been studied[14], the extent to which data collection forms impact cognition in clinical research data collection has not yet been investigated. Furthermore, cognitive science models and methodology have yet to be applied to medical record abstraction in clinical research or other secondary data use settings.

From cognitive science we know that distribution of information across internal and external representations, *i.e.,* in the user's mind and in the world, affects human task performance[13]. Additionally, representation can extend human performance through external cognition[12, 13]. Thus, data collection forms may impact data accuracy through form representation that supports distributed, *i.e.,* external cognition, and through the number and extent of mental steps between the data source and the collection form.

We applied the distributed cognition framework[12] and adapted Gong's information search model[11] to medical record abstraction, and applied them through a representational analysis to perform a systematic evaluation of paper data collection forms to 1) identify the type and extent of internal cognition required in medical record abstraction, and 2) to characterize the extent of support for external cognition in data collection forms.

Background

Medical record abstraction entails the identification of required data in the medical record, transformations of that data, and recording the data onto data collection forms. While two representations, 1) the source medical record, and 2) the destination data collection form, may impact data accuracy, secondary data users usually cannot impact how data are represented in the medical record. However, secondary users can control the representation of their data collection forms, *e.g.,* data collection forms often employ form instructions, prompts, and structural or graphical elements to guide form completion[8, 15, 16]. However, the presence of these elements, their extent and format are inconsistent on data collection forms [15]. Since data collection forms are present during the abstraction, and within control of the secondary data users, there is reason to believe that they may provide a mechanism to decrease cognitive load by increasing

support for external cognition during the abstraction process.

In his 2006 work, Gong applied the theory of distributed representation to explore how information distribution between internal and external representations affects information search performance[11]. He demonstrated that search task performance improved with increasing amounts of information represented externally[11]. Further, this and other work, has shown that search task performance improves when the scales[17] between the task and the data representation match[11, 12].

Because medical record abstraction is both a search and a cognitively intense process, the Gong model has particular utility for exploring and characterizing the extent to which data collection forms support distributed cognition in medical record abstraction. As such, we adapted Gong's model to the task of medical abstraction (Figure 1). Document boxes were added for medical record and data collection form representation. Task boxes were added for both documentation and abstraction tasks. *Remember*, *transform*, and *transcribe* are shown at the sub-task level, clearly delineating them from the *search* sub-task. In addition, *localize* from Gong's model was considered a direct search sub-task, while *compare* and *calculate* were relocated to the transform sub-task along with additional transformations *interpret*, *map*, and *scale conversion, i.e.,* a transformation from one scale to another, *e.g.,* ratio representation in the medical record, and data collection form representation as an ordinal category. Importantly, all tasks presented opportunities for distributed cognition. Light grey boxes were added for completeness but are not evaluated here.

In medical record abstraction, information is represented both in the medical record and on the data collection form. Therefore, there are opportunities for mismatch between 1) the represented information and the representing medical record, 2) the representing medical record and the representing data collection form 3) the represented information and the representing data collection form. Moreover, the search, remember, transform, and transcribe tasks are performed internally and likely increase working memory load unless external cognition artifacts exist.

While the medical record may have artifacts that enable external cognition for search tasks, the remember, transform, and transcribe tasks are unique to each secondary data use. Therefore, we expect that medical record representation will not provide significant opportunities for external cognitive support for these tasks. Thus, we evaluated paper data collection form external cognition artifacts rather than any particular medical record representation.

In medical record abstraction, virtually every data element by definition has a search task. Each data element may or may not have form artifacts supporting external cognition. Further, for each data element, zero to multiple transform tasks may apply. Each transform task required for a data element may or may not have an external cognition artifact.

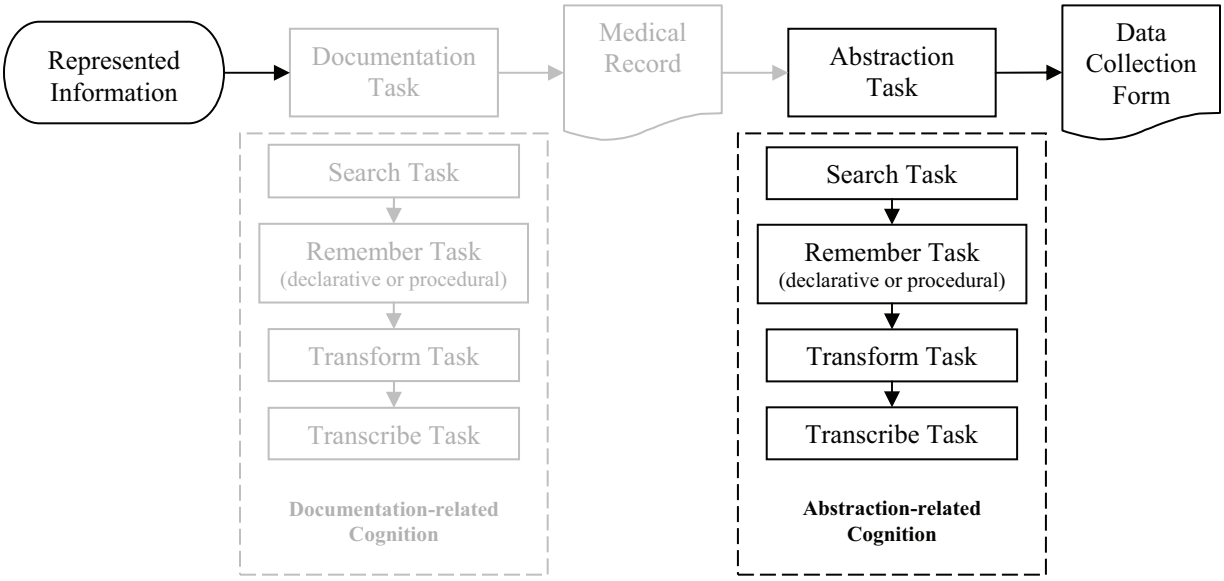


Figure 1. Model of Cognition in Medical Record Abstraction

Methods

Fifteen structured paper-based data collection form modules* were randomly selected from the data collection form library at the Duke Clinical Research Institute. The library houses data collection forms, many of which have been broken out by modules. We sampled the 256 available modules, randomly selecting 15 modules. Once nine unique trials were obtained, the remaining five modules were accepted sequentially only if they were from a trial already selected for the sample. This allowed comparison between forms within a trial.

The fifteen modules were from nine different clinical trials completed from 1992-2004. The module types and number of data elements per module are listed in Table 1. In total, 250 data elements were assessed.

Module Type	Data Elements*
Patient status (Trial 1)	25
Drug administration (Trial 1)	9
Canadian Cardiovascular Society Class (Trial 2)	4
Pacemaker mode change form (Trial 2)	32
Drug administration (Trial 3)	18
Platelet count (Trial 3)	12
Post procedure repeat catheterization (Trial 4)	18
Cardiac markers (Trial 5)	26
Clinical global impression (Trial 6)	6
Thyroid function tests (Trial 6)	12
Serum pregnancy test (Trial 6)	6
Medical history (Trial 7)	14
30 Day follow-up (Trial 8)	48
30 Day follow-up (Trial 9)	11
Cardiac enzymes (Trial 9)	9

* Only unique data elements were assessed and counted

Table 1. Characterization of Modules Selected for this Study.

Ten of the analyzed modules reflected different data content areas. Five of the analyzed modules were different representations (isomorphs) of the same content, *e.g.,* lab results, from different forms collected in different formats. The analysis of multiple instances of similar module content allowed assessment of differences in representation of similar data elements.

For the representational analysis, each data element was reviewed and classified by two independent reviewers (informatics graduate students in a health data display class) who were both novices to medical record abstraction. Our unit of analysis was the data element, *i.e.,* a form question and the associated response field† Each reviewer classified the following eight aspects of each data element: Scale of the represented information (nominal, ordinal, interval, ratio), Data collection form representing scale, Scale of the abstraction task, Presence of a search task (yes, no), Presence of external representation for the search task (yes, no), Type of other transform tasks, if present (compare, calculate, interpret, other), Enumeration of dimensions required to abstract the data element, and whether the Rule representation for the abstraction task and transformations was (internal, external).

Briefly, a search task is the locating of a data value, *e.g.,* finding documentation of the gender of a subject in the medical record. A transform task is a manipulation or conversion of a data value, *e.g.,* categorizing a specific medication according to class, or converting units on a lab value or drug dose. The rule is the logic that defines such a transformation. For this analysis, scale conversions were counted separately from other transformations.

A third person experienced in medical record abstraction adjudicated the work of the two independent reviewers; discrepancies were resolved by the adjudicator and final data were reviewed by all three reviewers.

We recognize that the representation in the medical record likely impacts cognition during medical record abstraction. However, we did not assess representation in the medical record because 1) medical record systems should optimize cognitive support for care delivery and clinical documentation rather than secondary data use, and 2) medical record representation differs from institution to institution. The impact of medical record representation on accuracy of abstracted data remains an area for future research. Similarly, abstractor experience may also impact cognitive demands by obviating of the need for transform through direct association, *e.g.,* body mass index of 47 equating to obese. Impact of abstractor experience remains an area for future research.

Results

Of the 250 data elements assessed, 98 (39%) were direct transcription, *i.e.,* once the value was located in the medical record, it could be copied directly onto the data collection form without transformation. For example, a blood pressure value recorded in the

* A module is a section of a data collection form containing data grouped by topicality, e.g., vital signs, physical exam, lab results. Modules are usually, but not always less than a page.

† Data element is formally defined in ISO/IEC 11179-1.

medical record in the same units as those required on the data collection form did not need interpretation or calculation if collected as a numeric value.

The majority of the data elements, 152 (61%) required transformation of some type. Cognitively, transformation means that a rule is required to change the data value from its source state to the destination state on the data collection form. Collection of age on the data collection form is an example; age would need to be calculated from the date of birth and the date of the screening visit. The types of transformation required include comparison, calculation, interpretation and mapping, shown by percentage in Table 2. Scale transformations were counted separately. In addition, 37 (15%) of the data elements required more than one transformation.

Transformation Type	Percent
Comparison	43%
Mapping (categorization)	29%
Interpretation (also included synthesis)	14%
Calculation	14%

Table 2. Characterization of Transformation

The data collection form representation for each data element was assessed and categorized as either supporting external cognition or not. As expected, external cognition for the 98 direct transcription data elements was supported by the data collection form. For these data elements, the form prompt and field structure made the search and transcription tasks perceptually evident, *i.e.,* no additional cognition on the part of the human abstractor required.

Supporting external cognition for the transformation (rule based) tasks, is more difficult. Unfortunately, the cognitively more complex data elements, *i.e.,* the 152 data elements requiring transformations, were not supported by form-based external cognition artifacts. One hundred and thirteen (74%) of these complex data elements, required internal cognition.

The number of dimensions, *i.e.,* individual distinct pieces of information, for example, using two dimensions, today's date and birth date to calculate today's age, required for each transformation was also assessed. The mean number of dimensions required for abstracting the data elements that needed a transformation was 2.6, with a range of 1 to 45 dimensions required. For example, an abstractor must compare, map, interpret, or calculate 45 distinct pieces of information. The highest number, 45 was a data element asking for enumeration of inclusion and exclusion criterion that were not met requiring internal assessment of each. Unless external cognitive support is provided, *e.g.,* a worksheet, or created by the abstractor, *e.g.*, a scratch pad, the values for each dimension are held in the abstractor's head prior to and during the transformation. Therefore, the dimension counts indicate the cognitive load required for the transformation.

Scale mismatch between the represented information, the abstraction task, and the data collection form representation further impacted internal cognitive demands on the abstractor by requiring mental transformations from one scale to another. Each data element was categorized three ways according to Steven's[17] nominal, ordinal, interval, and ratio scales, 1) the scale of the represented information, 2) the scale of the abstraction task, *i.e.,* the transformation or transcription, and 3) the scale of the data collection form representation. Table 3 shows the overall shift in scale from the represented information to the data collection form representation.

	Data Collection Form Representation			
Represented	Nominal	Ordinal	Interval	Ratio
Nominal (138)	138	0	0	0
Ordinal (19)	*16*	3	0	0
Interval (29)	*1*	*0*	28	0
Ratio (64)	*23*	*3*	*0*	38
	178	6	28	38

Table 3. Scale "down shift" from Represented Information to Data Collection Form

Overall, 43 (17%) of the data elements were reduced from the represented information scale to the data collection form representation. This down shift requires transformation, usually in the form of mapping, interpretation, or categorization. Thus, scale mismatch adds to the already significant cognitive load on the human abstractor.

Discussion

Although from only a limited evaluation in a small sample of paper data collection forms, the results reported here document the significant cognitive demands in medical record abstraction. A human can hold on average from 5-7 chunks of information in working memory[18]. Our results show that on average, the cognitive demands for many CRF data elements bump up against the limits of human cognition. Further, the 9% of data elements requiring four or more dimensions, when added to cognitive load from

transformations, can easily exceed working memory limits. Moreover, the paper data collection forms analyzed had few, if any, external cognition artifacts to support the most cognitively demanding data elements.

Other authors have cited requiring "abstractor judgment" or "interpretation" as a cause of errors in medical record abstraction[6, 8, 9, 10]. However, none have suggested why these errors occur or the nature of their relationship to other types of data error in medical record abstraction. Likewise, the literature does not suggest concrete methods of mitigating or preventing the resulting data errors. Our results contribute a possible explanation and mechanism for a portion of the data accuracy problem in medical record abstraction. In addition, the theory of distributed representation and the associated representational analysis used here can be applied to analyze data element representation on data collection forms and abstraction tasks to prevent or lessen the likelihood of cognitive limit related abstraction errors. The representational analysis used here can be performed during form design to identify high cognitive load data elements and to evaluate isomorphs. Such an application would improve form design by identifying high cognitive load data elements so that they can be replaced with lower cognitive load isomorphs. Further, both electronic medical records and electronic data collection forms, provide additional opportunities for decreasing cognitive load. Confirming these results in a larger and more diverse sample of forms and medical records, and evaluation of data accuracy from data collection form isomorphs are key next steps in this area of inquiry.

Conclusion

The cognitive load required for abstraction of 61% of the data elements in our sample was both high and unsupported with external cognition artifacts on the data collection forms, exceedingly so for 9% of the data elements. The high working memory demands are a possible explanation for the association of data errors in medical record abstraction with data elements that require abstractor interpretation, comparison, mapping or calculation. Existing methods of representational analysis can be applied to identify data elements with high cognitive demands and help form designers identify and avoid them in form design. Further, representational analysis provides a tool to analyze form isomorphs and identify those with the lowest cognitive demands.

The project described was supported by Grant Number UL1RR024128to Duke University and grant number UL1 RR 024148 to University of Texas Health Science Center Houston, from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH.

References

1. Gardiner RC. Quality considerations in medical records abstracting systems. J Med Syst. 1978;2(1):31-43.
2. Herrmann N, Cayten CG, Senior J, Staroscik R, Walsh S, Woll M. Interobserver and intraobserver reliability in the collection of emergency medical services data. Health Serv Res. 1980;15(2):127-143.
3. Robinson L, Hughes P. Use a streamlined approach for medical record abstraction. Qual Lett Healthc Lead. 1998;10(6):14-15.
4. Kerr EA, Smith DM, Hogan MM, Krein SL, Pogach L, Hofer TP, et al. Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures. Jt Comm J Qual Improv. 2002;28(10):555-565.
5. Nahm M, Johnson CM, Johnson T, Fendt K, Zhang J. Clinical research data quality literature review and systematic analysis. Clin Trials J. [submitted].
6. Allison JJ, Wall TC, Spettell CM, Calhoun J, Fargason CA Jr, Kobylinski RW, et al. The art and science of chart review. Jt Comm J Qual Improv. 2000;26(3):115-136.
7. Findley TW, Daum MC. Research in physical medicine and rehabilitation. III. The chart review or how to use clinical data for exploratory retrospective studies. Am J Phys Med Rehabil. 1991;70(1):S23-S30.
8. Banks NJ. Designing medical record abstraction forms. Int J Qual Health Care. 1998;10(2):163-167.
9. Beard CM, Yunginger JW, Reed CE, O'Connell EJ, Silverstein MD. Interobserver variability in medical record review: an epidemiological study of asthma. J Clin Epidemiol. 1992;45(9):1013-1020.
10. Feinstein AR, Pritchett JA, Schimpff CR. The epidemiology of cancer therapy. IV. The extraction of data from medical records. Arch Intern Med. 1969;123(5):571-590.
11. Gong, Y, Zhang, J. Effects of information displays for hyperlipidemia. HCI International 2009, in "Human Interface and the Management of Information. Information and Interaction" LNCS-5618, Springer, 2009(9) 503-512.
12. Zhang J, Norman DA. Representations in distributed cognitive tasks. Cog Sci. 1994;81(1):87-122.
13. Zhang J. A representational analysis of relational information displays. Int J of Human-Computer Studies. 1996;45:59-74.
14. Bang M, Timpka T. Cognitive tools in medical teamwork: the spatial arrangement of patient records. Methods Inf Med. 2003;42(4):331-336.
15. Spiker B, Schoenfelder J. Data collection forms in clinical trials. New York: Raven Press; 1991.
16. Society for Clinical Data Management. Good clinical data management practices [document on the Internet]. Milwaukee: SCDM; 2009 [cited 2009 Nov 4]. Available from: http://www.scdm.org.
17. Stevens, SS. On the theory of scales and measurement. Science. 1946;103(2684):677-680.
18. Miller GA. The magical number seven, plus or minus two. The Psychological Review. 1956;63(2):81-97