Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Error Rates of Data Processing Methods in Clinical Research: A Systematic Review and Meta-Analysis

Maryam Y. Garza (mygarza@uams.edu)
University of Arkansas for Medical Sciences
Tremaine Williams
University of Arkansas for Medical Sciences
Songthip Ounpraseuth
University of Arkansas for Medical Sciences
Zhuopei Hu
University of Arkansas for Medical Sciences
Jeannette Lee
University of Arkansas for Medical Sciences
Jessica Snowden
University of Arkansas for Medical Sciences
Anita C. Walden
University of Colorado Denver, Anschutz Medical Campus
Alan E. Simon
National Institutes of Health
Lori A. Devlin
University of Louisville
Leslie W. Young
University of Vermont
Meredith N. Zozus
The University of Texas Health Science Center at San Antonio

Research Article

Keywords: medical record abstraction, data quality, clinical research, clinical data management, data collection

Posted Date: January 20th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2386986/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background: Over the last 30 years, empirical assessments of data accuracy in clinical research have been reported in the literature. Although there have been articles summarizing results reported in multiple papers, there has been little synthesis of these results. Further, although notable exceptions exist, little evidence has been obtained regarding the relative accuracy of different data processing methods.

Methods: A systematic review of the literature was performed to identify clinical research studies that evaluated the quality of data obtained from data processing methods typically used in clinical research (e.g., medical record abstraction, optical scanning, single-data entry, and double-data entry). A total of 93 papers meeting our inclusion criteria were categorized according to their data processing methods. Quantitative information on data accuracy was abstracted from the articles and pooled. Meta-analysis of single proportions based on an inverse variance method and generalized linear mixed model approach of studies from the literature were used to derive an overall estimate of error rates across data processing methods for comparison.

Results: Review of the literature indicated that the accuracy associated with data processing methods varies widely, with error rates ranging from 2 errors per 10,000 fields to 2,784errors per 10,000 fields. The medical record abstraction process for data acquisition in clinical research was associated with both high and highly variable error rates, with a variability of 3 orders of magnitude in accuracy (70 – 2,784 errors per 10,000 fields). Error rates for data processed with optical methods were comparable to data processed using single-data entry (2 – 358 vs. 4 – 650 per 10,000 fields, respectively). In comparison, double-data entry was associated with the lowest error rates (4 – 33 per 10,000 fields).

Conclusions: Data processing and cleaning methods may explain a significant amount of the variability in data accuracy.

Background

Authors in the clinical research arena lament the scarcity of published information regarding data quality.[1–13] 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 and unnecessary.[14,15] This perception has resulted in minimal investigation and an associated small 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 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, such as electronic data capture (EDC), electronic patient reported outcomes (ePRO) systems, mobile health (mHealth), and direct electronic health record-to-EDC (EHR-to-EDC) technologies, understanding the quality of data from different available methods has become even more important.[16] 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.[17–20] Common data processing methods identified in the literature include: (1) chart review and abstraction versus direct electronic acquisition from electronic medical records (i.e., both types of medical record abstraction, or MRA); (2) use of vended or commercial 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 with central processing; and (4) single- versus double-data entry (with or without programmed edit checks). Data cleaning methods also vary greatly, from use of reports to identify irregularities in the data, to on-screen checks during data entry (e.g., programmed edit checks), to post-entry batch data processing. We define the 4 major processing methods considered in this review (MRA, optical scanning, single- vs. double-data entry with or without programmed edit checks) in Table 1.

Table 1
Definitions of Data Processing Methods in Clinical Research

Data Processing Method	Definition
Medical Record Abstraction (MRA)	A data processing method that involves the review and abstraction of data from patient records, often referred to as *chart review* or *chart abstraction*. Traditional MRA is a manual process which may or may not involve paper forms. More recently, use of data standards is modernizing data collection to allow for use of semi-automated approaches via direct EHR-to-EDC, electronic MRA.
Optical Scanning (OCR/OMR)	A data processing method that relies on software packages to "recognize characters from paper forms or faxed images, and these data are placed directly into the database."[21] Examples include *optical character recognition* (OCR) and *optical mark recognition* (OMR).
Single-Data Entry (SDE)*	Single-data entry involves a single person who collects data from the source (paper or electronic) and enters it into the study data capture system (e.g., EDC). SDE can be implemented with and without programmed edit checks (or on-screen data checks)*.
Double-Data Entry (DDE)*	Double-data entry involves 2 people who independently collect and enter data from the source to the study data capture system with a third, independent adjudicator to review and resolves any discrepancies. DDE can be implemented with and without programmed edit checks (or on-screen data checks)*.
**Programmed Edit Checks*	Programmed edit check – also referred to as *discrepancy checks*, *edit checks*, *on-screen checks* (OSCs), or *query rules* – are electronic data quality checks that are programmed into the study data collection system and are triggered by data entry, either in real-time as data is entered or in batches.

**Note. Both single- and double-data entry methods can be conducted with or without programmed edit checks.*

Complicating comparisons of different data processing methods is the significant variability in quantitative methods for assessing data accuracy across clinical research and other secondary data uses.[21–23] Data accuracy has often been measured in terms of database error rates, although registries also often assess percent completeness as well. The error rate is defined in the Society for Clinical Data

Management's (SCDM) Good Clinical Data Management Practices (GCDMP) document as the number of errors divided by the number of data values inspected.[21,24]

$$Error\ rate = \frac{number\ of\ errors}{number\ of\ values\ \ inspected} \quad (1)$$

As described in the GCDMP,[21] there are significant differences in the way errors and values are inspected and counted across different clinical research studies, even across studies conducted within the same institution. Based on these counting differences, the error rates obtained can differ by a factor of 2 or more.[21,23] 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. Here, we undertook a review and secondary analysis of the relevant literature to characterize data collection and processing methods utilized in clinical studies and registries. The objective of this study was to calculate and compare error rates across the various data processing methods described in the pooled literature.

Methods

Literature Review

A PubMed search on the Medical Subject Heading (MeSH) terms "data quality" AND (registry OR "clinical research" OR "clinical trial") was conducted to identify relevant citations (see Additional File 1, Appendix A, Item A1 for the full PubMed Search Strategy and Table A2 for the PRISMA Checklist). An initial search of the literature identified 350 citations. Duplicates were excluded, and the abstracts of the initial list of citations were screened for relevance against the eligibility criteria, resulting in 54 manuscripts. A search using PubMed related links and secondary and tertiary references was also conducted, identifying an additional 70 manuscripts, and yielding 124 manuscripts for full-text review. Through the full-text review, we identified the final set of 93 manuscripts (see Additional File 1, Appendix A, Reference List A3 and Table A4) included in the pooled literature analysis (Fig. 1).

Criteria for Manuscript Inclusion

Manuscripts were included in the analysis if: (1) they were published in peer reviewed journals indexed for retrieval or referenced by such and were publicly available; (2) they had a focus on secondary data use (e.g., clinical research, quality improvement, surveillance, research registries); (3) the database error rate was presented or resolvable (e.g., via number of errors identified and number of fields inspected, or contained sufficient information to calculate); (4) they described how the data were processed (e.g., MRA, optical scanning, single- or double-data entry); (5) they were written in the English language; and (6) the manuscript was the primary source for the error rate. The following parameters were optional: data

cleaning method (i.e., batch data cleaning), location of data processing (central data center vs. local healthcare facility), gold standard used, and scope of method of comparison.

Three types of data were collected from each article: (1) information about how data were processed; (2) information about how data quality was measured; and (3) the number of errors and number of fields inspected. Prior to quantitative data analysis, factors identified from items (1) and (2) were developed in a qualitative, iterative manner during the review of the articles. Attributes of the data processing and quality measurement methods reported were noted as each article was read. Natural groupings were organized into categories. These categories were later explored in the analysis to ascertain which (if any) of the factors might affect data quality.

Quantitative data accuracy information was abstracted from the manuscripts and pooled. Manuscripts were categorized by type of secondary data use, data processing method, and data accuracy assessment, representing combinations of common factors that are not mutually exclusive; thus, some articles appear in more than 1 category (see Additional File 1, Appendix A, Table A5). Information on the number of errors identified and the number of fields inspected was collected for each manuscript. We abstracted the number of errors reported and the total number of data fields (values) 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 the GCDMP.[21] In cases where the authors presented only normalized error rates, such as errors per 10,000 fields, the normalized denominator was assumed for the total number of fields inspected. For example, if the normalized error rate presented was 100 per 10,000 fields, we assumed 100 to be the total number of errors (numerator) and 10,000 to be the total number of fields (denominator). Where error rates for more than 1 database were provided in a manuscript, each individual assessment was included in this analysis. Where error rates for multiple data processing steps were provided, we included them; thus, the number of data points in this analysis is higher than the number of reviewed papers. Meta-analysis of single proportions[25,26] based on an inverse variance method[27] and generalized linear mixed model approach[28] of studies from the literature were used to derive an overall estimate of error rates across data processing methods for comparison. We also performed subgroup analyses where the data allowed.

For consistency, 1 rater was used to abstract the error rate information from the manuscripts. A sample of the manuscripts included in the analysis, comprising 10% of the total (standard for the domain), was re-evaluated by the primary rater 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 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 were comparable to those in a similar review paper of errors in EHRs.[29]

Statistical Analysis

All statistical tests were performed at a two-sided significant level of 0.05, and all analyses were carried out using the R package 'metafor' and 'meta'.[30,31] For each of the data processing methods, we used a standard meta-analytical method to calculate the pooled effect size and corresponding 95% confidence interval (CI). A meta-regression was implemented to compare the pooled effect among data processing methods. The degree of heterogeneity between studies were examined based on the Q-statistic and Higgins and Thompson's I^2 statistic. The I^2 statistic can be interpreted approximately as $\leq$ 25%, indicating low heterogeneity; 25% – 75% indicating moderate heterogeneity; and > 75%, indicating considerable heterogeneity.[32] The Q-statistic is typically underpowered for detecting true heterogeneity when the number of studies is small; therefore, we pooled data using a random effects model. The inter-study variance was evaluated by computing tau-squared $\left(\tau^2\right)$, which provides the estimated standard deviation of the underlying effects across studies.

Results

The data obtained from the comprehensive literature review are displayed in Fig. 2. Four manuscripts[33–36] presented only normalized error rates as errors per 10,000 fields; for these, the denominator was assumed for the total number of fields inspected. Each manuscript described a data quality assessment of 1 or more databases. Likewise, in some manuscripts, error rates were reported for more than 1 process step; for example, medical record-to-case report form (CRF) or source-to-CRF, CRF-to-first entry, first entry-to-second entry, or CRF-to-clean file. A total of 22 manuscripts reported results for more than 1 processing step, process, or database,[9,22,33,35,37–54] providing a total of 124 data points normalized as number of errors per 10,000 fields and demonstrating the dispersion over time of the health-related research literature with respect to data accuracy. During the meta-analysis, 9 outliers were detected and excluded (Fig. 2).

Database error rates ranged from 2–2,784 errors per 10,000 fields (having excluded outliers) across 4 data processing methods: MRA, optical scanning, single-data entry, and double-data entry. This 3 orders-of-magnitude range necessitated a logarithmic display. There appeared to be no pattern in the year-to-year reporting.

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 Table 2 and Fig. 3. Source-to-database and source-to-CRF (case report form) were combined into a single category labeled abstraction, or MRA, based on: (1) some of the manuscripts 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 2 processes being similar. Data in Fig. 3 have been horizontally jittered to more fully display the data set. The median (horizontal bar) is overlaid on each group. Forest plots were generated

for each of the 4 data processing methods and are available in the Appendices (see Additional File 2, Appendix B, Figures B1-B4).

Table 2
Pooled Effect Size of Error Rates (errors per 10,000) by Data Processing Method

Data Processing Method	Range	Pooled Proportion (95% CI)	p-value*
Abstraction (MRA)	70–2,784	657 (551, 772)	Ref
Optical Scanning (OCR/OMR)	2–358	74 (21, 160)	< 0.0001
Single-Data Entry (SDE)	4–650	29 (24, 35)	< 0.0001
Double-Data Entry (DDE)	4–33	14 (8, 20)	< 0.0001

* *Note. p-value were calculated through meta-regression.*

In exploring subgroups of the four main data processing methods, there is insufficient information in the literature about the MRA methods employed to further investigate possible causes for the variability in a subgroup analysis. For optical methods, there were too few data points to support a subgroup analysis. For single- and double-data entry, a review of the literature surfaced different variations on key entry, including single-data entry (1 person enters the data), single-data entry with on-screen data checks (1 person enters the data within a system employing programmatic, on-screen data checks or OSCs), and double-data entry (2 people independently enter data with a third, independent adjudicator to review and resolve discrepancies). Further variations on single-data entry found in the literature included, use of batch data cleaning, and the location of data processing. These results are provided in Additional File 3, Appendix C, Table C1. Due to the importance of this particular model, manuscripts reporting data accuracy from similar data processing configurations (e.g., central versus distributed data processing in the presence of OSC), were examined (see Additional File 3, Appendix C, Table C2). Sixty-eight studies (across 49 manuscripts) versus 49 studies (across 39 manuscripts) reported central versus distributed processing; while 7 studies (across 5 manuscripts) did not report the location of data processing (noted in Table A4, see Additional File 1, Appendix A).

Discussion

Medical Record Abstraction

Ordered by the median, MRA was associated with the highest error rate. Importantly, abstraction was also associated with significant variability. Notably, the error rates reported for MRA methods span 3 orders of magnitude, with error rates ranging from 70–2,784 errors per 10,000 fields. These results support claims that MRA, which remains the dominant method of data collection in retrospective and prospective research, is the most significant source of error in clinical research across data processing methods.[8,55]

Optical Scanning

Although optical scanning methods such as OCR and OMR have been touted as a faster, higher-quality or less resource-intensive substitute for manual data entry,[14,46,56–62] others have reported error rates with optical methods that were 3 times higher than manual keyboard data entry.[63] Based on the pooled literature, we found optical scanning error rates ranged from 2–358 errors per 10,000 fields. Optical methods were associated with a variability of 2 orders of magnitude in accuracy. Such variability may be influenced by: (1) the presence and type of data cleaning employed in processing the optical scans; (2) use of post-entry visual verification or pre-entry manual review; (3) training of form completers on handwriting; (4) differences in form compatibility with the software; (5) software configuration (e.g., recognition engine); and (6) variations in data quality assessment methods.

Single- vs. Double-Data Entry

Overall, single-entry error rates ranged from 4–650 errors per 10,000 fields, and double-entry error rates ranged from 4–33 errors per 10,000 fields. The subgroup analysis of single-data 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 EHRs, and data collected directly from patients via hand-held devices. Due to the problem of "alert fatigue," however, OSC may not be feasible in EHRs, where clinical alerts will often be a higher priority. The question of alert fatigue in these systems is an important topic for further research.

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, OSCs 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 2 categories labeled "All" in Appendix C, Table C1). The trend is the same for data entered at clinical sites ("Distributed, On Screen Checks" in Appendix C, Table C1), where OSCs were associated with a decreased average error rate of 28 errors per 10,000 fields, comparable to the average quality obtained from double-data entry. Importantly, the data presented in Appendix C, Table C1 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).

The predominant method of clinical research data collection in the near future will likely be web-based EDC.[64–68] 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 (Appendix Table B2).

Limitations

This study was 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 several limitations inherent in our work. First, as with any literature review, there is the possibility that we may

have missed relevant manuscripts. A lack of standard terminology for data processing methods potentially affected this analysis through high likelihood that relevant manuscripts were not identified. 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.[21,23] 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. Third, most of the manuscripts 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.

Future Direction

As data (increasingly captured electronically) are used to support clinical research, the effects of data quality on decision-making need thorough exploration. Potential effects of system usability and data processing methods on data quality should also be characterized to guide data management and planning choices. The variability and the magnitude of error rates reported in the literature should encourage additional evaluation of the impact of new technology and processes on data accuracy and subsequent decisions regarding whether the accuracy of the data is acceptable for the intended use.

Conclusion

Based on the pooled analysis of error rates from the published literature, we conclude that the MRA error rate is the highest and most variable compared to other clinical research data collection and processing methods, and that the observed error rates in the top quartile (904–2,784 errors per 10,000 fields) were high enough to potentially impact the results and interpretation of many clinical studies. Our findings suggest that reporting the results of a clinical study without specifying (1) the error rate, (2) the uncertainty in the error rate, and (3) the method used to measure the error rate limits the ability to interpret study findings.

List Of Terms & Abbreviations

Terms / Abbreviation	Meaning / Definition
MRA	Medical Record Abstraction
QC	Quality Control
EHR	Electronic Health Record
EDC	Electronic Data Capture
ePRO	Electronic Patient Reported Outcomes
mHealth	Mobile Health
eSource	The direct use of EHR (electronic source)[69] data in clinical research through the use of direct EHR-to-EDC technologies. The term 'eSource' is often used colloquially to refer to the method of direct capture, collection, and storage of electronic source data (e.g., EHRs, electronic patient diaries, or wearable devices) in an effort to streamline clinical research.[70,71]
Optical Scanning	A data processing method used in clinical research that relies on software packages to "recognize characters from paper forms or faxed images, and these data are placed directly into the database."[21]
OCR	Optical Character Recognition; an example of optical scanning
OMR	Optical Mark Recognition; an example of optical scanning
Single-Data Entry	A data processing method in which a single person enters data from the source (paper or electronic health record) to the study data capture system. In clinical research, single-data entry can be implemented with and without on-screen data checks. When on-screen checks are employed, a series of programmatic edit checks are actively running during data entry and will "fire" when a discrepancy is identified during data entry. The data entry person is then able to review and address discrepancies during data entry.
Programmed Edit Checks	A data processing method during which electronic data quality checks are programmed into the study data collection system and are triggered by data entry, either in real-time as data is entered field-by-field or upon the form being saved, or in batch based on some pre-determined criteria. Programmed Edit Checks are also referred to as *Discrepancy Checks*, *Edit Checks*, *On-Screen Checks*, or *Query Rules*.
OSC	On-Screen Checks
Double-Data Entry	A data processing method in which 2 independent persons enter data from the source (paper or electronic health record) to the study data capture system. Once data has been entered by both persons, a third, independent adjudicator reviews and resolves any discrepancies, resulting in a final, clean dataset in the study data capture system.
SCDM	Society for Clinical Data Management
NIH	National Institutes of Health
GCDMP	Good Clinical Data Management Practices

Terms / Abbreviation	Meaning / Definition
MeSH	Medical Subject Heading
CRF	Case Report Form
SDV	Source Data Verification
FDA	US Food and Drug Administration

Declarations

Ethics Approval and Consent to Participate

The literature review and meta-analysis received a determination of not human subjects research as defined in 45 CFR 46.102 by the UAMS IRB (IRB#239826). All methods were carried out in accordance with relevant guidelines and regulations.

Consent for Publication

Not Applicable

Availability of Data and Materials

The dataset(s) supporting the conclusions of this article is(are) included within the article (and its additional file(s)).

Competing Interests

The authors declare that they have no competing interests.

Funding

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award numbers UL1TR003107 and KL2TR003108, and by the IDeA States Pediatric Clinical Trials Network of the National Institutes of Health under award numbers U24OD024957, UG1OD024954, and UG1OD024955. The content is solely the responsibility of the authors and does not represent the official views of the NIH.

Authors' Contributions

MYG and MNZ conceived and designed the study. ACW, AES, LAD, LWY, JL, JS, and SO contributed significantly to the conception and design of the project. MYG and MNZ contributed significantly to the acquisition of the data. MYG led and was responsible for the data management, review, and analysis. AES, JL, MNZ, SO, TW, and ZH contributed significantly to the analysis and interpretation of the data. MYG led the development and writing of the manuscript. All authors reviewed the manuscript and contributed to revisions. All authors reviewed and interpreted the results and read and approved the final version.

Acknowledgments

We thank Phyllis Nader, BSE for her assistance with this project.

References

1. Knatterud GL, Rockhold FW, George SL, et al. Guidelines for quality assurance in multicenter trials: a position paper. *Control Clin Trials*. 1998;19(5):477-493. doi:10.1016/s0197-2456(98)00033-6
2. Division of Health Sciences Policy, Institute of Medicine (IOM). *Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making: Workshop Report*. (Davis JR, Nolan VP, Woodcock J, Estabrook RW, eds.). The National Academies Press; 1999. doi:10.17226/9623
3. 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(1):27-37. doi:10.1016/0197-2456(86)90005-x
4. 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(2):96-100. doi:10.1016/0197-2456(87)90034-1
5. Knatterud GL. Methods of quality control and of continuous audit procedures for controlled clinical trials. *Control Clin Trials*. 1981;1(4):327-332. doi:10.1016/0197-2456(81)90036-2
6. Freeland B. Moving toward continuous quality improvement. *J Intraven Nurs*. 1992;15:278-282.
7. 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(22):1469-1475. doi:10.1056/NEJM199511303332206
8. Blumenstein BA. Verifying keyed medical research data. *Stat Med*. 1993;12(17):1535-1542. doi:10.1002/sim.4780121702
9. 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(1-2):115-123; discussion 124. doi:10.1002/sim.4780090118
10. Day S, Fayers P, Harvey D. Double data entry: what value, what price? *Control Clin Trials*. 1998;19(1):15-24. doi:10.1016/s0197-2456(97)00096-2

11. 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 Oxf Engl 1990*. 1993;29A(7):943-947. doi:10.1016/s0959-8049(05)80199-6
12. 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 Lond Engl*. 2007;4(5):572-582. doi:10.1177/1740774507084106
13. Hall HI, Gerlach KA, Miller DS. Methods of quality management: Accuracy. *J Regist Manag*. 2002;29:72-77.
14. Nies MA, Hein L. Teleform: a blessing or burden? *Public Health Nurs Boston Mass*. 2000;17(2):143-145. doi:10.1046/j.1525-1446.2000.00143.x
15. Crewson PE, Applegate KE. Fundamentals of Clinical Research for Radiologists: Data Collection in Radiology Research. *Am J Roentgenol*. 2001;177(4):755-761. doi:10.2214/ajr.177.4.1770755
16. Nahm ML, Pieper C, M Cunningham M. Quantifying Data Quality for Clinical Trials Using Electronic Data Capture. *PloS One*. 2008;3:e3049. doi:10.1371/journal.pone.0003049
17. Garza MY, Williams TB, Myneni S, et al. Measuring and Controlling Medical Record Abstraction (MRA) Error Rates in an Observational Study. Published online March 11, 2022. doi:10.21203/rs.3.rs-1225727/v1
18. Helms RW. A distributed flat file strategy for managing research data. In: *Proceedings of the ACM 1980 Annual Conference*. ACM '80. Association for Computing Machinery; 1980:279-285. doi:10.1145/800176.809982
19. Collen MF. Clinical research databases--a historical review. *J Med Syst*. 1990;14(6):323-344. doi:10.1007/BF00996713
20. Nahm M. Clinical Data Management: Why we do what we do? *Data Basics*. Published online 2007:13.
21. SCDM. Good Clinical Data Management Practices (GCDMP). Published online 2013. https://scdm.org/wp-content/uploads/2019/10/21117-Full-GCDMP-Oct-2013.pdf
22. Arts DGT, 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 JAMIA*. 2002;9(6):600-611. doi:10.1197/jamia.m1087
23. Nahm M, Dziem G, Fendt K, Freeman L, Masi J, Ponce Z. Data quality survey results. *Data Basics*. 2004;10:13-19.
24. Caballero I, Verbo E, Calero C, Piattini M. MMPRO: Amethodology based on ISO/IEC 15939 to draw up data quality measurement processes. In: *Proceedings of the 13th International Conference on Information Quality, MIT*. ; 2008:340.
25. Barker TH, Migliavaca CB, Stein C, et al. Conducting proportional meta-analysis in different types of systematic reviews: a guide for synthesisers of evidence. *BMC Med Res Methodol*. 2021;21(1):189. doi:10.1186/s12874-021-01381-z

26. Schwarzer G, Rücker G. Meta-Analysis of Proportions. *Methods Mol Biol Clifton NJ*. 2022;2345:159-172. doi:10.1007/978-1-0716-1566-9_10
27. Lee CH, Cook S, Lee JS, Han B. Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. *Genomics Inform*. 2016;14(4):173-180. doi:10.5808/GI.2016.14.4.173
28. McCulloch CE. AN INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS. *Conf Appl Stat Agric*. Published online April 28, 1996. doi:10.4148/2475-7772.1314
29. Hogan WR, Wagner MM. Accuracy of Data in Computer-based Patient Records. *J Am Med Inform Assoc*. 1997;4(5):342-355.
30. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health*. 2019;22(4):153-160. doi:10.1136/ebmental-2019-300117
31. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *J Stat Softw*. 2010;36:1-48. doi:10.18637/jss.v036.i03
32. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560. doi:10.1136/bmj.327.7414.557
33. 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(3 Suppl):84S-94S. doi:10.1016/0197-2456(89)90044-5
34. Fabri J, Beller E. An assessment of the value of double data entry and interactive checking [abstract]. *Control Clin Trials*. 1993;14:405.
35. 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 Off J Eur Soc Vasc Surg*. 1997;14(6):468-474. doi:10.1016/s1078-5884(97)80126-8
36. 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(1):101-107. doi:10.1177/009286159903300113
37. 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 JAMIA*. 2008;15(3):386-389. doi:10.1197/jamia.M2381
38. Crombie IK, Irving JM. An investigation of data entry methods with a personal computer. *Comput Biomed Res Int J*. 1986;19(6):543-550. doi:10.1016/0010-4809(86)90028-5
39. Birkett NJ. Computer-aided personal interviewing. A new technique for data collection in epidemiologic surveys. *Am J Epidemiol*. 1988;127(3):684-690. doi:10.1093/oxfordjournals.aje.a114843
40. Jasperse DM, Ahmed SW. The Mid-Atlantic Oncology Program's comparison of two data collection methods. *Control Clin Trials*. 1989;10(3):282-289. doi:10.1016/0197-2456(89)90068-8

41. 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(1):67-75. doi:10.1093/ije/18.1.67
42. Reynolds-Haertle RA, McBride R. Single vs. double data entry in CAST. *Control Clin Trials*. 1992;13(6):487-494. doi:10.1016/0197-2456(92)90205-e
43. 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(6):482-488. doi:10.1016/0197-2456(94)90005-1
44. Horbar JD, Leahy KA. An assessment of data quality in the Vermont-Oxford Trials Network database. *Control Clin Trials*. 1995;16(1):51-61. doi:10.1016/0197-2456(94)00019-y
45. 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.
46. 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(1):39-49. doi:10.1016/s0197-2456(97)00094-9
47. Goldhill DR, Sumner A. APACHE II, data accuracy and outcome prediction. *Anaesthesia*. 1998;53(10):937-943. doi:10.1046/j.1365-2044.1998.00534.x
48. Velikova G, Wright EP, Smith A, et al. Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. *J Clin Oncol Off J Am Soc Clin Oncol*. Published online 1999. doi:10.1200/JCO.1999.17.3.998
49. Håkansson I, Lundström M, Stenevi U, Ehinger B. Data reliability and structure in the Swedish National Cataract Register. *Acta Ophthalmol Scand*. 2001;79(5):518-523. doi:10.1034/j.1600-0420.2001.790519.x
50. 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(5):656-659. doi:10.1007/s00134-002-1272-z
51. 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(5):560-569. doi:10.1016/s0197-2456(03)00089-8
52. Rostami R. Data quality in a global environment. In: Society for Clinical Data Management; 2004.
53. 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(9):918-923. doi:10.1016/j.jclinepi.2005.02.004
54. 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 JABFM*. 2007;20(2):151-159. doi:10.3122/jabfm.2007.02.060069

55. Jansen ACM, van Aalst-Cohen ES, Hutten BA, Büller HR, Kastelein JJP, Prins MH. Guidelines were developed for data collection from medical records for use in retrospective analyses. *J Clin Epidemiol*. 2005;58(3):269-274. doi:10.1016/j.jclinepi.2004.07.006
56. Smyth ET, Emmerson AM. Survey of infection in hospitals: use of an automated data entry system. *J Hosp Infect*. 1996;34(2):87-97. doi:10.1016/s0195-6701(96)90133-4
57. Jørgensen CK, Karlsmose B. Validation of automated forms processing. A comparison of Teleform with manual data entry. *Comput Biol Med*. 1998;28(6):659-667. doi:10.1016/s0010-4825(98)00038-9
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(5):471-484. doi:10.1038/sj.jea.7500043
59. Dyck PJ, Turner DW, Davies JL, et al. Electronic case-report forms of symptoms and impairments of peripheral neuropathy. *Can J Neurol Sci J Can Sci Neurol*. 2002;29(3):258-266. doi:10.1017/s0317167100002043
60. 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(5):425-437. doi:10.1016/s0010-4825(03)00012-x
61. Quan KH, Vigano A, Fainsinger RL. Evaluation of a data collection tool (TELEform) for palliative care research. *J Palliat Med*. 2003;6(3):401-408. doi:10.1089/109662103322144718
62. Hardin JM, Woodby LL, Crawford MA, Windsor RA, Miller TM. Data collection in a multisite project: Teleform. *Public Health Nurs Boston Mass*. 2005;22(4):366-370. doi:10.1111/j.0737-1209.2005.220410.x
63. 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 Symp Comput Appl Med Care*. Published online 1995:86-90.
64. Weber BA, Yarandi H, Rowe MA, Weber JP. A comparison study: paper-based versus web-based data collection and management. *Appl Nurs Res ANR*. 2005;18(3):182-185. doi:10.1016/j.apnr.2004.11.003
65. Bruland P, Doods J, Brix T, Dugas M, Storck M. Connecting healthcare and clinical research: Workflow optimizations through seamless integration of EHR, pseudonymization services and EDC systems. *Int J Med Inf*. 2018;119:103-108. doi:10.1016/j.ijmedinf.2018.09.007
66. Garza M, Myneni S, Nordo A, et al. eSource for Standardized Health Information Exchange in Clinical Research: A Systematic Review. *Stud Health Technol Inform*. 2019;257:115-124.
67. Garza M, Myneni S, Fenton SH, Zozus MN. eSource for Standardized Health Information Exchange in Clinical Research: A Systematic Review of Progress in the Last Year. *J Soc Clin Data Manag*. 2021;1(2). doi:10.47912/jscdm.66
68. Garza MY, Prior F, Sanford JA, Sexton KW, Zozus MN. Innovative solutions to streamline data collection, exchange, and utilization in translational research. *J Clin Transl Sci*. 2022;6(1):94. doi:10.1017/cts.2022.277

69. FDA. Electronic Source Data in Clinical Investigations: Guidance for Industry. Published online September 2013. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/electronic-source-data-clinical-investigations
70. Garza MY, Rutherford M, Myneni S, et al. Evaluating the Coverage of the HL7 ® FHIR ® Standard to Support eSource Data Exchange Implementations for use in Multi-Site Clinical Research Studies. *AMIA Annu Symp Proc AMIA Symp*. 2020;2020:472-481.
71. Garza MY, Rutherford MW, Adagarla B, et al. Evaluating Site-Level Implementations of the HL7 FHIR Standard to Support eSource Data Exchange in Clinical Research. *Stud Health Technol Inform*. 2021;281:397-401. doi:10.3233/SHTI210188

Figures

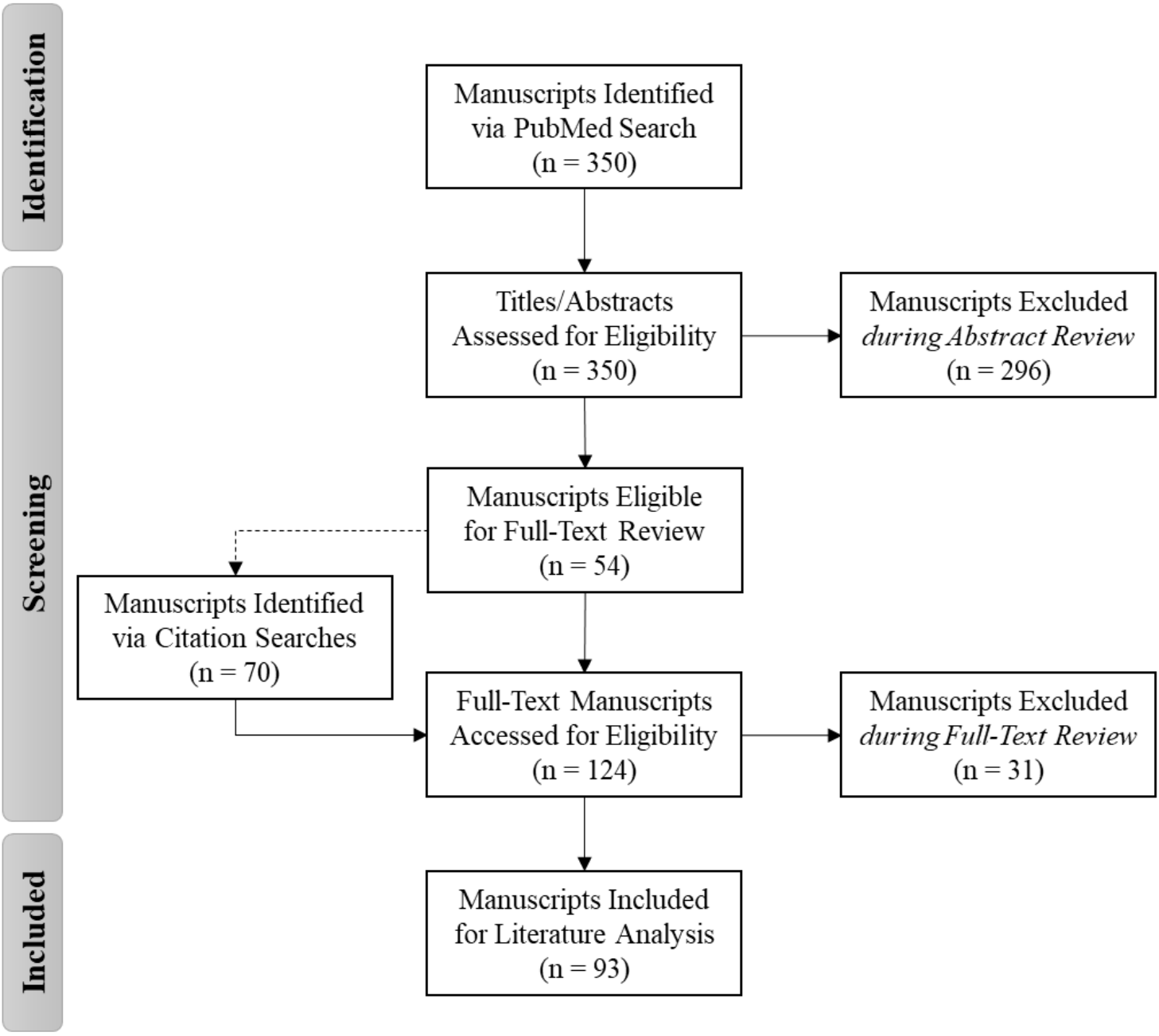

Figure 1

PRISMA Diagram: Identification of Data Quality Literature for Pooled Analysis

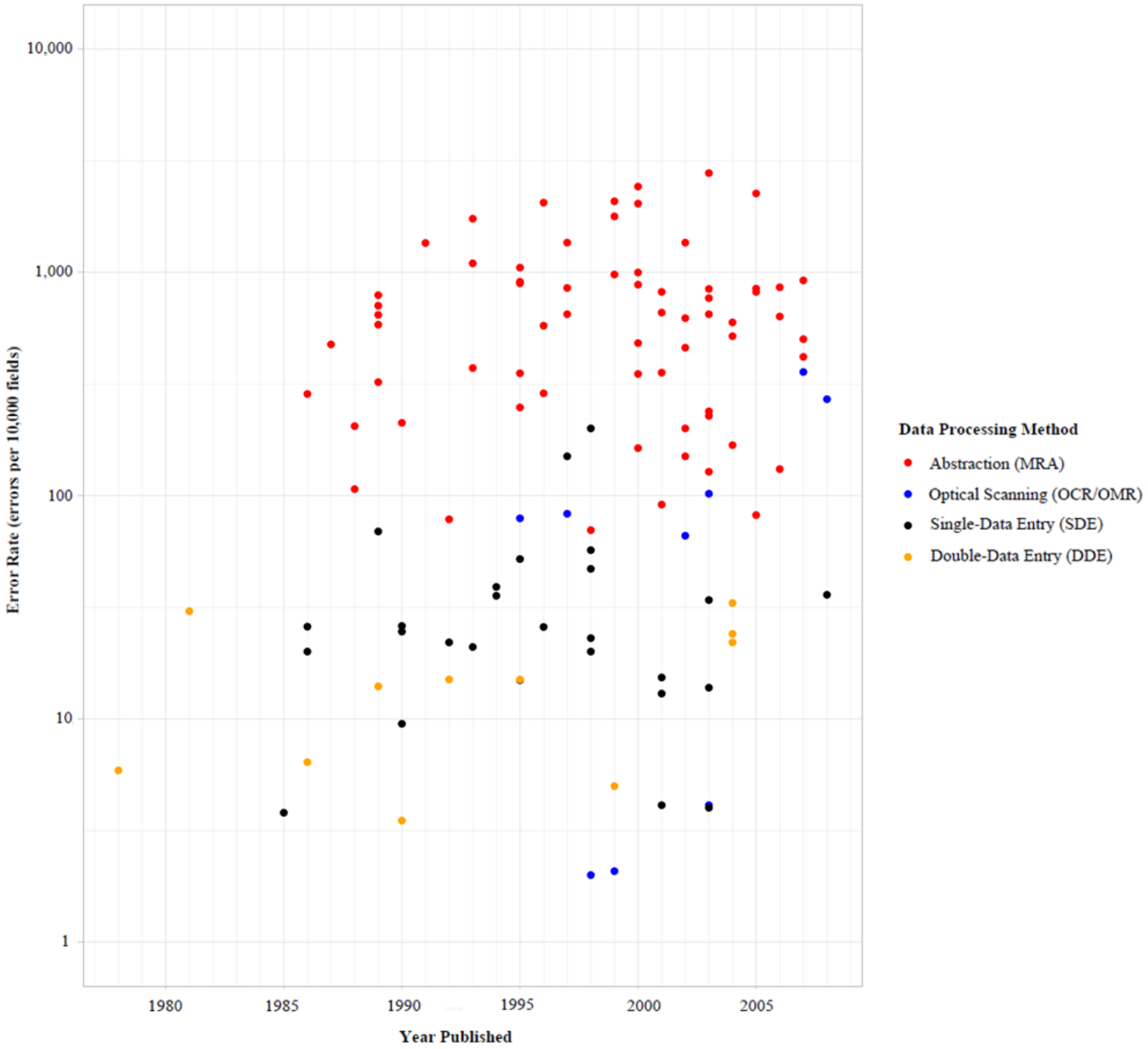

Figure 2

Chronological Survey of the Database Error Rate Literature

Note. 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.

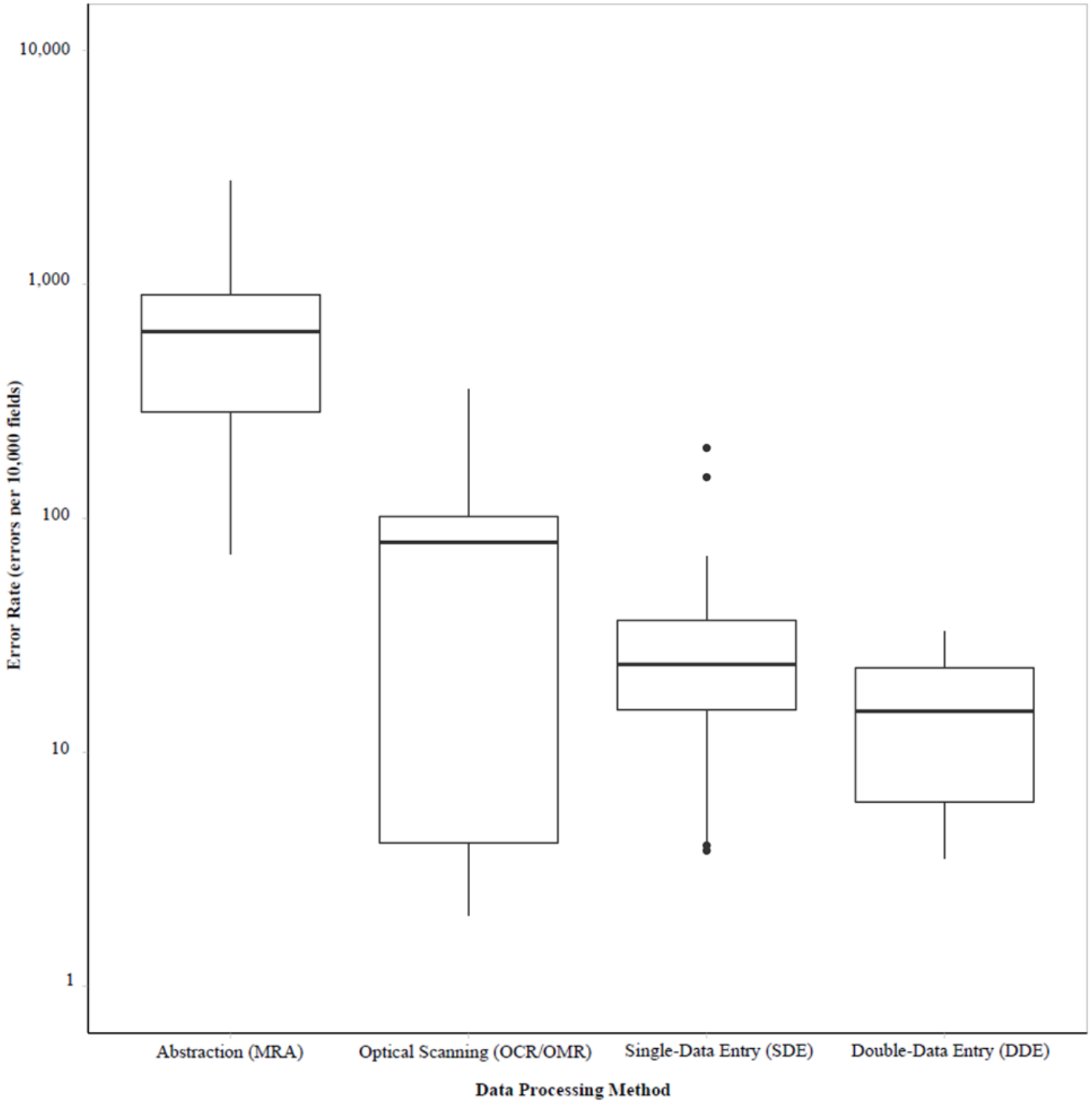

Figure 3

Data Error Rates for Different Processing Methods

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1AppendixA.docx
- AdditionalFile2AppendixB.docx
- AdditionalFile3AppendixC.docx