

Editorial

The journey to transparency, reproducibility, and replicability

Suzanne Bakken

Regardless of the type of biomedical and health informatics research conducted (eg computational, randomized controlled trials, qualitative, mixed methods), transparency, reproducibility, and replicability are crucial to scientific rigor, open science, and advancing the knowledge base of our field and its application across practice domains. These principles are also essential to high-quality publications in *Journal of the American Medical Informatics Association (JAMIA)*. Transparency is reflected by explicit, clear, and open communication about the methods and procedures used to obtain the research results and is foundational to reproducibility (ability to repeatedly obtain the same results from data) and replicability (ability of other investigators to observe the same result under identical conditions).^{1,2} In the following paragraphs, I summarize key strategies from a number of authors^{1–3} as well as my own thoughts in 4 categories (data, code, connect, publish) and, when applicable, describe their relationship to publishing in *JAMIA*. While the principles apply across types of research, the relevance of some strategies varies.

DATA

- Record how each result was produced.³ The analysis workflow, including pre- and postprocessing steps, should be presented in your *JAMIA* manuscript as a figure or in an online supplement.
- Record all intermediate results.³
- Deposit data in a repository. *JAMIA* authors can deposit their data in Dryad free of charge (<https://datadryad.org/journal/1067-5027>). Dryad provides a basic level of curation, assignment of a digital object identifier, and long-term data storage. *JAMIA* articles that include deposited data will receive additional promotion by both the American Medical Informatics Association and Oxford University Press.
- Publish a dataset paper if the dataset is unique and available for reuse by others. For *JAMIA*, a dataset paper is considered a Research and Applications paper.

CODE

- Avoid manual data manipulation steps.³
- Record and, if possible, archive the exact versions of all external programs used.³ This includes not only data analysis programs,

but also algorithms used to extract, filter, and reduce multidimensionality of the data at the various stages of the analysis workflow. These programs should be reported in your *JAMIA* manuscript.

- Apply and record version-control strategies to all custom scripts (eg, Git).³
- Record underlying random seeds and which analysis steps involve randomness for analyses that include randomness.³ This should be reported in your *JAMIA* manuscript.

CONNECT

- Store raw data behind and the code used to create plot or figures.³
- Generate layered output for inspection in levels of detail (eg, use HTML file hypertext links to go from summary to detail).³ Layered output can be included in the online supplement for your *JAMIA* manuscript.
- Connect textual statements to underlying results.³ Such annotations are a typical part of qualitative research and supported by multiple software packages (eg, ATLAS.ti, Dedoose, NVivo). Quantitative packages include the Sweave function in R that creates dynamic reports for integration into LaTeX or LyX documents. Literate programming approaches combine analysis code, plots, and text narrative that can be shared: for example, Jupyter Notebooks (for R, Python, and Julia; <http://jupyter.org>), R Markdown (<http://rmarkdown.rstudio.com>), and matlabweb (<https://www.ctan.org/pkg/matlabweb>).² The site for your literate programming content should be referenced in your *JAMIA* manuscript.

PUBLISH

- Register your research protocol.^{1,4} While registration of research protocols for clinical trials is routine at ClinicalTrials.gov, other registries are emerging. For example, PROSPERO (<https://www.crd.york.ac.uk/prospere/>) is an international register for

prospective systematic reviews. Your registered protocol should be noted in your *JAMIA* manuscript.

- Make a draft version of a manuscript available on a preprint server such as *arXiv* (<https://arxiv.org/>) or *bioRxiv* (<https://www.biorxiv.org/>). This does not preclude publication of the manuscript in *JAMIA*.
- Attend to general guidelines and checklists for publishing research and *JAMIA* author guidelines when preparing a *JAMIA* manuscript. General guidelines vary by research design and purpose, and include CONSORT (Consolidated Standards of Reporting Trials), (Strengthening the Reporting of Observational Studies in Epidemiology), PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), and SRQR (Standards for Reporting Qualitative Research). Many guidelines are available on the EQUATOR (Enhancing the Quality and Transparency of Health Research) Network website (<https://www.equator-network.org/reporting-guidelines/>).
- Provide public access to scripts, runs, and results.^{1,3} Multiple generic and health research-specific platforms support public access, and include GitHub (<https://github.com/>); CIELO (Collaborative Informatics Environment for Learning on Health Outcomes),⁵ which is now part of the Clinical and Translational Science Award Clinical Data to Health program (<https://ctsa.ncats.nih.gov/cd2h/cd2h-labs/>); and literate programming packages (eg, Jupyter Notebook, R Markdown). The site relevant to the contents of your *JAMIA* manuscript should be referenced.
- Clearly disclose research outcomes as primary, secondary, or exploratory in experimental research.¹ Such specification is foundational to distinguish hypothesis testing from hypothesis-generating analyses. For *JAMIA*, an online supplement provides the opportunity to balance manuscript length with full disclosure.

Each highlighted paper in this issue illustrates more than 1 aspect of transparency, reproducibility, or replicability. In terms of transparency, Price et al⁶ chronicled the 25-year history of implementing national information systems to support the National Health Service in England through an exploratory, retrospective longitudinal case study that examined National Health Service structural reorganizations alongside concurrent national information technology strategies. Guided by strong structuration theory, they reviewed strategic plans, legislation, and health policy documents, and constructed schemata for evolving structure and strategy to create a conceptual model. Their findings suggest that a long-term health information technology strategy may be impeded by volatility of the implementation environment as organizational structures and relationships change. While the authors point out the need for further research on the structure-strategy dyad, the lessons from this highly visible megaproject have high relevance to others.

Gonul et al⁷ describe the development and validation of a template-based digital intervention design framework to support just-in-time adaptable interventions. Their automated approach provides the foundation for transparency, reproducibility, and replicability by (1) enabling experts to explicitly specify decision points, intervention options, tailoring variables and decision rules (through a rule definition language), and proximal or distal outcomes; and (2) dynamically tailoring intervention delivery strategies with respect to timing, frequency, and type (content) of interventions based on a personalization algorithm. The authors provide evidence for the extensibility of their design as well as preliminary validation of the personalization algorithm.

Mercaldo et al⁸ report the study design used for the eMERGE (Electronic Medical Records and Genomic) Network's survey of perspectives on broad consent and data sharing in biomedical research. To ensure that understudied populations were adequately represented (eg, minorities, those from rural areas, those with low educational attainment), they combined electronic health record data and U.S. Census data to construct sampling strata by imputing missing electronic health record data using the most frequent (mode) value from the patient's census block group. Reproducibility and replicability are enabled by provision of details about the algorithmic approach as well as an example from 1 study site. Moreover, the authors are transparent about the limitations of their approach.

To complement existing approaches for screening, brief intervention, and referral to treatment programs at trauma centers that have proven efficacy for reducing alcohol consumption and decreasing injury recurrence, Afshar et al⁹ applied the clinical Text Analysis and Knowledge Extraction System (cTAKES) natural language processor and machine learning to clinical emergency department notes for trauma admissions using the as the reference standard. Regarding transparency, the authors specified hypotheses about the expected performance of cTAKES on the corpus and developed the rule-based keyword algorithm that in advance of the cTAKES results. To support reproducibility and replicability, the reference standard (Alcohol Use Disorders Identification Test) and various programs or frameworks used are described, the supplement lists the Unified Medical Language System semantic types used for alcoholic beverages, and the source code is publicly available in Apache cTAKES SVN (subversion) repository.

Lu et al,¹⁰ from the Lister Hill National Center for Biomedical Communications at the National Library of Medicine, developed a novel spell-checking tool, CSpell, that handles nonword errors, real-word errors, word boundary infractions, punctuation errors, and combinations of these errors and infractions for consumer-generated questions. Their approach uses dual embedding within Word2vec for context-dependent corrections in combination with dictionary-based corrections in a 2-stage ranking system. They also developed various splitters and handlers to correct word boundary infractions. The dual-embedding model shows a significant improvement in F1 score compared with the general practice of using cosine similarity with word vectors in Word2vec for context ranking and the 2-stage ranking system shows an almost 5% improvement in F1 score compared with the best 1-stage ranking system. In support of reproducibility and replicability, the software and the CSpell test set are available at <https://umlslex.nlm.nih.gov/cSpell>.

The scientific community in general, our informatics community, and *JAMIA* are on a journey toward increased transparency, reproducibility, and replicability. I ask the *JAMIA* readers and authors to engage with the editorial team to make our journey one that meets the needs of our biomedical and health informatics community while advancing an open science framework.

REFERENCES

1. Miguel E, Camerer C, Casey K, et al. Social science. Promoting transparency in social science research. *Science* 2014; 343 (6166): 30–1.
2. Gorgolewski KJ, Poldrack RA. A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS Biol* 2016; 14 (7): e1002506.
3. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. *PLoS Comput Biol* 2013; 9 (10): e1003285.

4. Shamseer L, Moher D, Clarke M, *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015; 350: g7647.
5. Payne P, Lele O, Johnson B, Holve E. Enabling open science for health research: collaborative informatics environment for learning on health outcomes (CIELO). *J Med Internet Res* 2017; 19 (7): e276.
6. Price C, Green W, Suhomlinova O. Twenty-five years of national health IT: exploring strategy, structure and systems in the English NHS. *J Am Med Inform Assoc* 2019; 26 (3): 188–97.
7. Gonul S, Namli T, Huisman S, Erturkmen G, Toroslu I, Cosar A. An expandable approach for design and personalization of digital, just-in-time adaptive interventions. *J Am Med Inform Assoc* 2019; 26 (3): 198–210.
8. Mercaldo N, Brothers K, Carrell D, *et al.* Enrichment sampling for a multi-site patient survey using electronic health records and census data. *J Am Med Inform Assoc* 2019; 26 (3): 219–27.
9. Afshar M, Phillips A, Karnik N, *et al.* Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *J Am Med Inform Assoc* 2019; 26 (3): 254–61.
10. Lu CJ, Aronson AR, Shooshan SE, Demner-Fushman D. Spell checker for consumer language (CSPELL). *J Am Med Inform Assoc* 2019; 26 (3): 211–8.