This is a "preproof" accepted article for *Journal of Clinical and Translational Science*.

This version may be subject to change during the production process.

10.1017/cts.2025.28

Data science and Artificial Intelligence in Biology, Health and Healthcare

Peter L. Elkin, MD[1], Christopher Lindsell, PhD[2], Julio Facelli, PhD[3], Manisha Desai, PhD[4], Chunhua Weng, PhD[5], Heidi Spratt, PhD[6], Shari Messinger, PhD[8], Lemuel Russell Waitman, PhD[9], JaMor Hairston, MSHI, MS[10], Ruth O'Hara, PhD[4] , Jareen Meinzen-Derr, PhD[7]

1. Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo
2. Duke University
3. Department of Biomedical Informatics, University of Utah
4. Stanford University School of Medicine
5. Department of Biomedical Informatics, Columbia University
6. Department of Biostatistics and Data Science, University of Texas Medical Branch
7. Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine
8. Department of Public Health Sciences, Division of Biostatistics and Bioinformatics, University of Miami
9. Department of Biomedical Informatics, Biostatistics and Medical Epidemiology, Missouri University School of Medicine
10. Department of Biomedical Informatics, Emory University School of Medicine

Corresponding Author: Peter L. Elkin, MD, MACP, FACMI, FNYAM, FAMIA, FIAHSI, UB Distinguished Professor and Chair, Department of Biomedical Informatics, Professor of Internal Medicine, Professor of Surgery, Professor of Pathology and Anatomical Sciences, Professor of Psychiatry, Professor of Orthopedics, 77 Goodell Street, Suite 540, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York

http://medicine.buffalo.edu/departments/biomedical-informatics.html

elkinp@buffalo.edu

Authors contributions:

Peter L.Elkin, MD[1] – wrote the article, edited and reviewed.

Christopher Lindsell, PhD[2] – wrote the article, edited and reviewed.

Julio Facilli, PhD[2] - edited and reviewed.

Manisha Desai, PhD - edited and reviewed.

Chunhua Weng, MD, PhD - edited and reviewed.

Heidi Spratt, PhD - edited and reviewed

Shari Messinger, PhD - edited and reviewed

Russ Waitman, PhD - edited and reviewed

Jamor Hairston, PhD - edited and reviewed

Ruth O'Hara, MD – edited and reviewed

Jareen Meinzen-Derr, PhD – authored and reviewed.

Acknowledgements (Funding):

This work has been supported in part by grants from NIH NLM T15LM012495, R25LM014213, NIAAA R21AA026954, R33AA0226954 and NCATS UL1TR001412 (PLE). We acknowledge resources/support from the Miami Clinical and Translational Science Institute, which is supported by the National Center for Advancing Translational Sciences, National Institutes of Health, Award Number UM1TR004556(SM), Research reported in this publication was supported by the Washington University Institute of Clinical and Translational Sciences grant UL1TR002345 from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) (RW). This publication was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number UL1TR001873 (CW). This publication was supported by the National Center for Advancing Translational Sciences, National Institutes of Health CTSA award to the Utah Clinical and Translational Science Institute UM1TR004409 (JF). Research reported in this publication was supported by the Stanford University Clinical and Translational Sciences Award UL1TR003142 from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) (MD). This publication was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number UL1TR001425(JMD).

COI: None

https://doi.org/10.1017/cts.2025.28 Published online by Cambridge University Press

Biomedical science is at an inflection point where classical thinking about how to conduct translational research is undergoing a transformation. Enabled by now available data sources combined with exponential advances in AI, Statistics and Computer Science we can achieve advances that require levels of complexity that we could not have previously thought possible. [1] To evaluate modern biomedical questions of significance requires data scientists to serve as partners in data driven science, as well as continuing their leadership in data science, biomedical informatics, and biostatistics research.[2] The future is hobbled by an insufficient data science workforce cross-trained in the biomedical sciences. As resources shift to centralized and connected repositories of increasingly complex data, the data science and AI community urges immediate action to ensure we draw unbiased conclusions from the vast information available to modern researchers.[3] Specifically,

- In order for science to be data driven, qualified data scientists should be integrally engaged from the project onset to limit fatal flaws in methodology, analysis and model development.
- Major data science initiatives should be coordinated to prevent redundancy and inefficiencies. This includes the consistent use of common accepted standards for codifying health data.
- It is critical to rapidly grow a qualified workforce of data scientists who can be counted on to exhibit a common core set of competencies, as well as to provide quantitative science training in the biomedical science curricula.
- Artificial Intelligence (AI) in biomedicine is a rapidly evolving field with significant promise and also significant risk. Attention to the elimination of bias, Ethical Legal and Social Issues (ELSI), the selection of data and populations on which to train models, and other issues must continue to be addressed.
- As we learn how to safely and effectively train and use AI in healthcare and research applications, it is essential that we establish frameworks for evaluating when and if data science methods and AI are fit for purpose.
- Emphasizing the FAIR principles (Findable, Accessible, Interoperable and Reusable, along with data cleaning, storage, indexing and data exchange and accessibility of data fit for purpose should be emphasized. Publishing high quality data resources should be rewarded similarly to publications in peer reviewed journals.

https://doi.org/10.1017/cts.2025.28 Published online by Cambridge University Press

- Congress should increase the NIH budget specifically earmarked for data science research and education, informatics and biostatistics to address this immediate and critical need. We suggest that good homes for the funding include NCATS, NLM and then other interested ICs.

1) Set expectations for the early inclusion of data scientists in all clinical biomedical research.

Data-scientists provide novel and independent contributions requiring a deep knowledge base and creativity. While team science is lauded, until data-scientists are recognized for their pivotal contributions as collaborating scientists on the team, there will continue to be a barrier for data-scientists to focus on addressing biomedical questions. Funding agencies can continue to help address this barrier by ensuring that data-centered science includes data scientists in named, recognized leadership roles, such as multi-PI roles, as is being increasingly observed.

2) Coordinate major data science and AI initiatives among federal agencies

There are increasing demands on the biomedical research community to meet the data science needs of multiple federal agencies. When multiple common data models are imposed, it creates competing workstreams for an already overburdened and under-supported health information technology ecosystem and for the data science workforce. Lack of consistency creates inefficiencies and sharing of non-interoperable data can lead to errors in reporting. Mapping between data standards, data repositories and common data models creates efficiency, but at the cost of information and quality. Deciding on a common set of interoperability standards would be the true path to semantic interoperability for our research and clinical care enterprise. Semantic Interoperability requires formalisms well beyond selecting a common data model.

3) Grow the data science and AI workforce and establish core competencies and subspecialties for data scientists in fields of biomedical research.

https://doi.org/10.1017/cts.2025.28 Published online by Cambridge University Press

There is a national shortage of data scientists to meet the needs of the future as envisioned. This deficit can be addressed both by creating new training programs and expanding existing programs. As established and emerging data sciences including artificial intelligence and machine learning evolve, recognition of subspecialty skills when forming scientific teams is critical - biostatisticians, informaticians, and data engineers are not interchangeable. The paradigm for training may also need to evolve: as programs sprint to keep pace with rapidly changing technology it is critical to ensure that the core competencies needed for extracting knowledge from data are addressed alongside the range of technical proficiencies that can be achieved. Doctoral, post-doctoral, and continuing data science training and education programs should consider the sheer breadth of the activities that constitute data science and ensure a common understanding of the general and specialty-specific knowledge base and competencies needed to train a functional data scientist workforce[4]. We would seek to create a generation of life-long learners.

Addressing these challenges will require a major coordinated response from academia, industry and government agencies. The many existing efforts can be optimized for impact with a clearer recognition and understanding of both core competencies and subspecialty skills, by applying standards for technology stacks and data stores, and by providing a reward and value system for data scientists contributing to biomedical research. Without a coordinated focus on these issues, we expect the gaps in expertise across the research enterprise to widen. However, greater attention to the biomedical data science workforce has the potential to catalyze efficiencies and shorten timelines to having robust answers to biomedical questions bringing new treatments to patients more rapidly. Strong data science is expected to result in not only high-quality reproducible research, but also a transformation of how clinical and translational research is performed. Specifically, we can use existing data to help us understand what studies are more likely to show positive results. We can judge likely toxicities ahead of expensive phase III clinical trials, and in so doing limit the number of negative trials. We can improve recruitment to and execution of clinical trials using real world evidence. We can more effectively do post-market surveillance[5] to find rare but serious side effects and deliver the right treatments to the right patients at the right time.

https://doi.org/10.1017/cts.2025.28 Published online by Cambridge University Press

References:

[1] Friedman AB, Delgado MK, Weissman GE. Artificial Intelligence for Emergency Care Triage—Much Promise, but Still Much to Learn. *JAMA Netw Open.* 2024;7(5):e248857. doi:10.1001/jamanetworkopen.2024.8857

[22] Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication. *JAMA Surg.* Published online June 05, 2024. doi:10.1001/jamasurg.2024.1621

[3] Zink A, Chernew ME, Neprash HT. How Should Medicare Pay for Artificial Intelligence? *JAMA Intern Med.* Published online May 28, 2024. doi:10.1001/jamainternmed.2024.1648

4 Resendez S, Franklin G, Stephens R, Maness H, Chamala S, Elkin PL. Analyzing the Efficacy of an Open Access Biomedical Informatics Boot Camp. Stud Health Technol Inform. 2024 Aug 22;316:1545-1546. doi: 10.3233/SHTI240711. PMID: 39176500.

5 Elkin PL, Mullin S, Mardekian J, Crowner C, Sakilay S, Sinha S, Brady G, Wright M, Nolen K, Trainer J, Koppel R, Schlegel D, Kaushik S, Zhao J, Song B, Anand E. Using Artificial Intelligence With Natural Language Processing to Combine Electronic Health Record's Structured and Free Text Data to Identify Nonvalvular Atrial Fibrillation to Decrease Strokes and Death: Evaluation and Case-Control Study. J Med Internet Res. 2021 Nov 9;23(11):e28946. doi: 10.2196/28946. PMID: 34751659; PMCID: PMC8663460.