Journal of Law, Medicine & Ethics (2026), **54**, 3, 420–422
doi:10.1017/jme.2026.10268

CAMBRIDGE UNIVERSITY PRESS

Columns: Currents in Contemporary Bioethics

Assuring Data Authenticity in the Age of Generative AI

Jeff J. H. Kim[1], Mohammad Hosseini[2], Haavi Morreim[3], C. K. Gunsalus[4], and David B. Resnik[5]

[1]University of Illinois Chicago, Chicago, Illinois, USA; [2]Northwestern University, Chicago, Illinois, USA; [3]Internal Medicine, College of Medicine, The University of Tennessee Health Science Center, Memphis, Tennessee, USA; [4]Coordinated Science Lab, University of Illinois Urbana-Champaign Grainger College of Engineering, Champaign, Illinois, USA and [5]National Institute of Environmental Health Sciences, Durham, North Carolina, USA

Abstract

This commentary examines the challenges that generative artificial intelligence poses to the authenticity and integrity of scientific data. It proposes data-attestation requirements to strengthen accountability, preserve access to original research records, and help journals and institutions distinguish authentic evidence from synthetic or fabricated data.

Keywords: Generative AI; Data Authenticity; Data Provenance; Research Misconduct; Research Integrity

About This Column
The Currents in Contemporary Bioethics column features examinations of recent developments and trends in bioethics. If you would like to submit to this column, please contact JLME editor-in-chief Aaron Kesselheim at akesselheim@bwh.harvard.edu.

Science is founded upon data that are reproducible, reliable, and authentic. However, generative artificial intelligence (GenAI) technologies are blurring the line between authentic data that are casually linked to real phenomena and synthetic data that are created by computer algorithms. The ability of individuals or organizations (such as paper mills) to use GenAI tools to produce realistic-looking data and misrepresent them as genuine threatens the integrity and trustworthiness of the research enterprise. Data attestation statements can play an important role in helping to protect science from the emerging threat posed by GenAI synthetic data and images.

GenAI's ability to mimic real data has advanced dramatically in the last few years.[1] GenAI tools can now produce realistic radiologic, fluorescence, and microscopy images; tabular data and data tables; electronic health records; and time-series records.[2] Studies have shown that radiologists have difficulty distinguishing between synthetic chest radiographs and genuine radiographic images and most people cannot tell the difference between AI-generated images of human faces and real images of faces.[3]

Since GenAI synthetic data have become highly realistic, the scientific community now faces a data authenticity crisis. Provenance, that is, the traceable history of how something came to be and reached its current state, is essential for distinguishing between authentic and synthetic data.[4] Authentic scientific data can be linked to real phenomena observed by human beings or recorded by machines.[5] Original research records provide evidence of data provenance and are therefore essential for determining whether data (or images) are real or have been fabricated or manipulated.[6] Sequestering research records (i.e., seizing them and keeping them safe) is one of the steps that an institution takes when it initiates a research misconduct inquiry. The purpose of sequestration is to preserve provenance by preventing original records from being tampered with or destroyed. It also serves as the trigger for the person accused of misconduct (the respondent) to state what happened to the original data. In addition to helping to assure data integrity, preservation of research records helps to promote other aspects of good science, including quality control, data sharing, collaboration, and replication of results.

Today, distinguishing between GenAI data and images and authentic ones will be increasingly difficult for institutions, funding organizations, or journals. Examination of the research records, especially the original data, usually plays a crucial role in determining whether fabrication or falsification has occurred. However, if the data or images have been produced by GenAI, there will be no records that an institution could use to authenticate them. When this happens, the respondent could claim that the records have been lost or destroyed. If loss or destruction of records is not treated as evidence of research misconduct, as is the case in current US Public Health Service regulations, the investigating institution could find itself in a difficult position when a researcher claims that they have lost the records.[7] Although losing research records may be reckless behavior, it does not constitute misconduct unless the institution can independently demonstrate, by forensic analysis or other means, that fabrication or falsification occurred.

Corresponding author: Jeff Kim; Email: jkim671@uic.edu
Cite this article: J.J.H. Kim, M. Hosseini, H. Morreim, C.K. Gunsalus, & D.B. Resnik. "Assuring Data Authenticity in the Age of Generative AI," *Journal of Law, Medicine & Ethics*, **54**, 3 (2026): 420–422. https://doi.org/10.1017/jme.2026.10268

© National Institutes of Health (NIH) and the Author(s), 2026. To the extent this is a work of the US Government, it is not subject to copyright protection within the United States. Published by Cambridge University Press on behalf of American Society of Law, Medicine & Ethics. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/jme.2026.10268 Published online by Cambridge University Press

Box 1: Sample Data Attestation Statement [for journal]

Data Possession and Retention

1. Our research team possesses or has authorized access to the original data (including image data) on which this research is based and
2. Our team will retain the data for at least ten years from the time of submitting this manuscript and
3. Take appropriate measures to keep it safe and secure.

Generative AI Use

4. If there are synthetic data in the manuscript (such as data produced by generative artificial intelligence), they are clearly marked as such and their scientific purpose is described and justified, for example, to correct for bias in the original data.

Data Sharing

5. Our research team will make data available upon request from the editors or reviewers for the purpose of evaluating our research.
6. If our manuscript is published, our team will share data (including copies of the original data) as widely as possible, to the extent that sharing does not violate important ethical values, such as protecting the privacy of human research participants or the welfare of communities.

Enforcement

7. Following publication, if our research team refuses reasonable requests to share data for the purposes of reproducing or validating our research, the journal may retract our article.

Thus, not treating failure to produce research records as evidence of misconduct creates a legal loophole that undermines the enforcement of research misconduct policies. The point here is not that every missing record indicates misconduct; records do sometimes get lost or corrupted. The point is that deceptive use of powerful GenAI tools, combined with weak legal consequences for absent records, may create a new form of plausible deniability.

Technical solutions, such as provenance frameworks and blockchain-based certification systems, can partially address this problem.[8] However, these tools have technical and practical limitations. It can be very costly to implement and maintain the information technology infrastructure required to operate these systems, making them infeasible for smaller institutions. More importantly, provenance systems only certify documented history, not the authenticity of the underlying data itself. If fabricated data are introduced before provenance capture begins, a provenance system may faithfully preserve the history of a fraud rather than prevent it.

A stronger response to the data authentication problem the scientific community now faces would be to combine technical provenance systems with enforceable attestations to the authenticity of the data or images. Attestations are used by most journals to ensure that submitted manuscripts are original, have not been previously published, and contain no material for which required copyright permissions have not been secured. We propose that in light of GenAI impacts on data integrity, scientific organizations, including journals, academic institutions, and research sponsors, should consider using attestation statements to help ensure that submitted data or images are authentic. To help promote research integrity, these statements could also include statements about use of GenAI, recordkeeping, and data sharing. (See Box 1 for an example of a data attestation statement.)

Some journals already require authors to submit original images and deposit data in publicly accessible websites. What is needed now is a more comprehensive accountability framework that ties publication or funding to a documented commitment that the underlying evidentiary record exists and remains available for review. In this way, data attestation would not replace provenance but would give it institutional oversight and legal force. Attestation shifts research integrity policy from post-hoc suspicion to prospective accountability. If concerns emerge about stand-alone published data or data associated with a manuscript, journals and institutions would no longer be left asking whether the necessary material ever existed. Authors would already have attested that it did and that they can provide it, so the burden of proof would fall on them to show that their data are authentic.

Such a framework would not eliminate misconduct, nor would it be a substitute for fair investigation of misconduct allegations. However, it would close a legal loophole in which plausibility and guile can override authenticity and accountability. The response to the data integrity issues created by GenAI tools should not be to reject them outright but to strengthen provenance standards for scientific data, require clearer stage-specific disclosure of GenAI use, and adopt submission requirements that preserve the link between published evidence and raw data. Without these links, scientific credibility becomes easier to simulate than to earn.

Data, code, and materials availability. No new data sets, code, or materials were generated for this article. All information supporting the conclusions of this article is contained in the manuscript.

Acknowledgments. The contributions of the NIH author (DBR) are considered works of the United States Government. The findings and conclusions in this paper are those of the authors and do not necessarily reflect the views of the NIH, the US Department of Health and Human Services, or the US Government.

Author contribution. Conceptualization: DBR; Supervision: DBR; Writing – original draft: JJHK, MH, HM, CKG, DBR; Writing – review & editing: JJHK, MH, HM, CKG, DBR.

Funding. National Institutes of Health Intramural Research Program (DBR); National Center for Advancing Translational Sciences grant UM1TR005121 (MH). National Institute on Alcohol Abuse and Alcoholism grant F30AA033278 (JJHK).

Competing interests. Authors declare that they have no competing interests.

Jeff J.H. Kim, MSE, is a MD/PhD Candidate at University of Illinois College of Medicine.

Mohammad Hosseini, PhD, is Assistant Professor of Preventative Medicine at Northwestern University Feinberg School of Medicine.

Haavi Morreim, JD, PhD, is Professor in the College of Medicine, University of Tennessee Health Science Center.

C. K. Gunsalus, JD, is Director of the National Center for Principled Leadership and Research Ethics at the University of Illinois Urbana-Champaign.

David B. Resnik, PhD, JD, is a bioethicist at the National Institute of Environmental Health Sciences and Senior Advisor for Research Integrity in the Office of Intramural Research at the National Institutes of Health.

References

1. David B. Resnik et al., "GenAI synthetic data create ethical challenges for scientists. Here's how to address them," *Proceedings of the National Academy of Sciences* 122, no. 9 (2025): e2409182122, https://doi.org/10.1073/pnas.2409182122; Isaac Kohane, "The H-Index of Suspicion: How Culture, Incentives, and AI Challenge Scientific Integrity," *NEJM AI* 3, no. 1 (2025): AIe2501273. https://doi.org/10.1056/AIe2501273.

https://doi.org/10.1017/jme.2026.10268 Published online by Cambridge University Press

2. Resnik et al., "GenAI synthetic data create ethical challenges for scientists. Here's how to address them"; Kohane, "The H-Index of Suspicion: How Culture, Incentives, and AI Challenge Scientific Integrity"; Jeff JH Kim, et al., "Generative AI can fabricate advanced scientific visualizations: Ethical implications and strategic mitigation framework," *AI and Ethics* 5, no. 5 (2025): 4481–4493, https://doi.org/10.1007/s43681-024-00439-0; Ho Young Park et al., "Realistic High-Resolution Body Computed Tomography Image Synthesis by Using Progressive Growing Generative Adversarial Network: Visual Turing Test," *JMIR Medical Informatics* 9, no. 3 (2021): e23328, https://doi.org/10.2196/23328; Edward Choi et al., "Generating multi-label discrete patient records using generative adversarial networks," in Proceedings of the 2nd Machine Learning for Healthcare Conference, *Proceedings of Machine Learning Research* 68 (2017): 286–305; Lei Xu et al., "Modeling Tabular Data Using Conditional GAN," in Proceedings of the 33rd International Conference on Neural Information Processing Systems, (Curran Associates Inc., 2019): 7335–7345; Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar, "Time-series Generative Adversarial Networks," in Proceedings of the 33rd International Conference on Neural Information Processing Systems, (Curran Associates Inc., 2019): 5508–5518; Miso Jang et al., "Image Turing test and its applications on synthetic chest radiographs by using the progressive growing generative adversarial network," *Scientific Reports* 13, no. 1 (2023): 2356, https://doi.org/10.1038/s41598-023-28175-1.
3. Elizabeth Gibney, "Hey ChatGPT, write me a fictional paper: these LLMs are willing to commit academic fraud," *Nature* 651, no. 8105 (2026): 286–287, https://doi.org/10.1038/d41586-026-00595-9; Sophie J. Nightingale and Hany Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences* 119, no. 8 (2022): e2120481119, https://doi.org/10.1073/pnas.212048111.
4. Resnik et al., "GenAI synthetic data create ethical challenges for scientists. Here's how to address them"; Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan, "Data provenance: Some basic issues," in *International Conference on Foundations of Software Technology and Theoretical Computer Science*, (Springer Berlin Heidelberg, 2000): 87–93.
5. Resnik et al., "GenAI synthetic data create ethical challenges for scientists. Here's how to address them."
6. Adil E. Shamoo and David B. Resnik, *Responsible Conduct of Research* (Oxford University Press, 2009).
7. Public Health Service Policies on Research Misconduct, 42 C.F.R. § 93.105(b)(1) (2024).
8. Paolo Missier, Khalid Belhajjame, and James Cheney, "The W3C PROV family of specifications for modelling provenance metadata," in *Proceedings of the 16th Iinternational Conference on Extending Database Technology*, (Association for Computing Machinery, 2013): 773–776, https://doi.org/10.1145/2452376.2452478; Xueping Liang et al., "Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability," in *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, (IEEE, 2017): 468–477, https://doi.org/10.1109/CCGRID.2017.8.

https://doi.org/10.1017/jme.2026.10268 Published online by Cambridge University Press