News and Perspectives

Data Governance Lessons From an Unvalidated Dataset

Cliff Dominy, JMIR Correspondent

Key Takeaways

- An open access dataset has highlighted how bad data can propagate through the research ecosystem.
- When trained on unvalidated datasets, machine learning can amplify misinformation, erode trust in science, and harm vulnerable populations.
- Enforced data provenance systems could play a key role in preventing bad data from corrupting the scientific record.

When an unvalidated dataset recently made it into the medical literature, it exposed several weaknesses in data governance. The dataset was uploaded to Kaggle—a large online platform where users can share publicly accessible data, code, and models—and was fundamentally flawed [1]. Its developer had compiled unverified images of children from websites related to autism to train an artificial intelligence (AI) model to "detect the presence of autism or the absence thereof" from the scraped images [2].

A sharp-eyed reviewer exposed the problem only at the publication stage; by December 2025, it was estimated that over 90 published papers had incorporated the bad data, leading to investigations and double-digit retractions [1,3].

These kinds of data integrity and governance failures are particularly consequential because of how early they occur in the research life cycle, and with open access datasets fueling large-scale machine learning and other AI research, analyses can be generated and published at unprecedented speed and scale, allowing data issues to propagate more rapidly throughout the research ecosystem. Far from an isolated incident [4-8], this situation highlights the need for more robust and proactive data governance solutions.

The Impact of Bad Data

Anne Borden is an autism advocate, journalist, and author of the upcoming book *The Informed Parent*—a decision-making guide for parents of autistic children. For Borden, the priority here is to learn from this "bizarre story" and fix the system without delay. "You really have to stop misinformation being perpetuated under the banner of science," she says, "because once it's out there, you're done. The Internet is forever."

Bad Data, Bad Science: Who Should Fix This?

Who are the custodians of good data during their migration from a spreadsheet to the scientific record? What role should each stakeholder play in maintaining data integrity? While responsibility for data governance is distributed across many actors (including researchers and regulators), data-sharing platforms, research and funding institutions, and academic publishers help determine how data are shared, vetted, and ultimately incorporated into the scientific record.

The Data-Sharing Platforms

Open access databases and data repositories, like Kaggle and GitHub, are popular resources that software developers and data scientists use to train their machine learning algorithms for free. Software development benefits from these repositories, yet the datasets they host often lack the documentation, governance, and quality practices required for careful medical research or clinical algorithm development [9].

Alan Katz, MBChB, MSc, CCFP, is a professor of family medicine and community health sciences and a senior scientist at the Manitoba Centre for Health Policy (MCHP). Katz found the dataset revelations "both shocking, but also not surprising" due to the rapid expansion of open access databases and their widespread use in machine learning and AI research. The Kaggle-style data-sharing platforms differ sharply from established medical databases, such as those maintained by the MCHP, which employs full-time staff tasked with validating all new data before uploading them. Katz says, "We take our ethical standards as seriously as clinical trials do."

Elizabeth Green, DPhil, is a lecturer in business and law at the University of the West of England, Bristol. Her research focuses on data integrity, and while she has seen cases like this before, she doesn't believe locking data away is necessarily the solution [10]. For example, DermAtlas—an open-source medical database of skin conditions—is a "fantastic resource," she says, and "extremely helpful, especially in [diagnosing] some extremely rare cases." To balance the risks and benefits of open data, the focus should instead be on building better governance systems.

The Institutions

Other stakeholders in the data transformation journey are the institutions that conduct primary medical research and the public agencies that fund that research. Is it time to adopt and enforce international data integrity and ethics standards at all research institutions, or would this be an affront to academic freedom?

Funding bodies have traditionally taken a dim view of researchers that waste public funds on bogus science, which impacts their future grants. Indeed, in many but not all regions of the world, funding is contingent on maintaining ethical research standards. In Canada, Katz says, “our existence is 100% dependent on having those strict ethical guidelines.”

The Journals

The research integrity pipeline involves several stakeholders, with each having distinct roles in maintaining the standards of academic research. Gatekeepers in the system—one of the last lines of defense—are the academic journals. Journals have a vested interest in maintaining high academic standards and may be well placed to dictate the terms of engagement.

Felix Ritchie, PhD—a colleague of Elizabeth Green—developed the Five Safes data integrity framework for just this purpose [11]. Ritchie describes it as “a flexible structure for thinking about [data],” which includes the provenance and ethics of data use. Numerous organizations worldwide have adopted the Five Safes framework to date, and Australia has recently legislated it [12].

Viewed through an ethical lens, the Five Safes could form the backbone of a data provenance system that requires compliance before a manuscript can be considered for publication.

Data Provenance: The Five Safes in Action

Ritchie’s Five Safes framework allows for effective data validation and, when combined with modern ethical standards, can restore trust by filtering data sources through five discrete tests:

1. Safe Project: Data should be ethically collected and clinically validated by experts.
2. Safe People: Researchers accessing the data must be qualified and specifically trained in using AI-based datasets.
3. Safe Data: Data should be independently validated, and any accesses or modifications should be tracked.
4. Safe Settings: Were health data acquired in a clinical setting and the data securely stored?
5. Safe Outputs: Were valid methodologies and statistics used to derive the results?

Restoring Data Integrity

How can one implement a data provenance system?

Ritchie feels that applying the Five Safes framework to an ethical dataset is the way forward. “There is a need for a register of validated, ethical datasets,” he says,” that would really be a game changer.”

A possible workflow could include the following:

1. Data are collected by medical experts and validated by a third-party certification service.
2. The data are stored in an accredited data registry and protected by blockchain cybersecurity—the same technology that safeguards financial transactions.
3. Researchers access these datasets and use them for approved research purposes.
4. A submitted manuscript would need ethical approval and a data security certificate before verification by a journal’s research integrity team.

Ritchie sums it up nicely: “Unless you use a validated data set, you’re not getting published, mate.” That’s a powerful incentive.

Opportunity for Self-Reflection and Correction

Machine learning and other AI technologies have the capacity to transform medical research in ways we are only beginning to understand. However, human frailties, such as blind trust in open access data and lack of institutional ethical oversight within our publish-or-perish culture, have shown how quickly such technologies can amplify misinformation.

While the impact of this situation was ultimately contained, it is nevertheless an important opportunity for self-reflection among all in the research ecosystem. It’s a chance and, perhaps, a responsibility to fix the flaws and prevent history from repeating itself.

Keywords: data integrity; data quality; data management; data provenance; data sharing; research integrity; scientific misconduct; retraction of publication; research ethics; artificial intelligence

Conflicts of Interest

None declared.

References

1. McMurray C. Exclusive: Springer Nature retracts, removes nearly 40 publications that trained neural networks on ‘bonkers’ dataset. The Transmitter. Dec 8, 2025. URL: https://www.thetransmitter.org/retraction/exclusive-springer-nature-retracts-removes-nearly-40-publications-that-trained-neural-networks-on-bonkers-dataset/ [Accessed 2026-03-09]
2. Info.txt. Google Drive. URL: https://drive.google.com/file/d/1zMQgyQvYiYyxx9J5jw3jrLGTS0p19Rep/view [Accessed 2026-03-09]
3. Expression of concern: data mining-based model for computer-aided diagnosis of autism and gelotophobia: mixed methods deep learning approach. JMIR Form Res. Jan 23, 2026;10:e91833. [doi: 10.2196/91833] [Medline: 41576282]

4. Toraih EA, ElWazir M, Elshazli RM, Hussein MH, Fawzy MS, Elroukh SM. Rapid publication during crises: analyzing retractions during the Covid-19 pandemic. Ethics Med Public Health. 2025;33:101136. [doi: 10.1016/j.jemep.2025.101136]
5. León FR. RETRACTED: likely electromagnetic foundations of gender inequality. Cross Cult Res. Apr 2023;57(2-3):239-263. [doi: 10.1177/10693971221143577]
6. Lancet, NEJM retract controversial COVID-19 studies based on Surgisphere data. Retraction Watch. Jun 4, 2020. URL: https://retractionwatch.com/2020/06/04/lancet-retracts-controversial-hydroxychloroquine-study/ [Accessed 2026-03-09]
7. Okyay RA, Kocyigit BF, Qumar AB, Yessirkepov M, Sumbul HE. Fifty years of retracted medical publications from 1975 to 2024: a comprehensive analysis of trends, reasons, and countries using the Retraction Watch database. J Korean Med Sci. Dec 1, 2025;40(46):e300. [doi: 10.3346/jkms.2025.40.e300] [Medline: 41327922]
8. Peng K, Mathur A, Narayanan A. Mitigating dataset harms requires stewardship: lessons from 1000 papers. arXiv. Preprint posted online on Aug 6, 2021. [doi: 10.48550/arXiv.2108.02922]
9. Avlona NR, Cheplygina V, Jiménez-Sánchez A, et al. Copycats: the many lives of a publicly available medical imaging dataset. Presented at: Advances in Neural Information Processing Systems 37; Dec 10-15, 2024:113383-113404; Vancouver, BC, Canada. [doi: 10.52202/079017-3602]
10. ripleywk. Case study: Dr Elizabeth Green -trust as foundation: enabling safe data access for public good. UWE Bristol blogs. Feb 28, 2025. URL: https://blogs.uwe.ac.uk/research-external-engagement/case-study-dr-elizabeth-green-trust-as-foundation-enabling-safe-data-access-for-public-good/ [Accessed 2026-03-09]
11. Desai T, Ritchie F, Welpton R. Five Safes: designing data access for research. Department of Accounting, Economics and Finance, Bristol Business School, University of the West of England, Bristol; 2016. URL: https://ideas.repec.org/p/uwe/wpaper/20161601.html [Accessed 2026-03-09]
12. Data Availability and Transparency Code 2022. Australian Government Office of the National Data Commissioner. 2022. URL: https://www.datacommissioner.gov.au/support/resources/data-availability-and-transparency-code-2022 [Accessed 2026-03-09]

Please cite as:
Dominy C
Data Governance Lessons From an Unvalidated Dataset
J Med Internet Res 2026;28:e94518
URL: https://www.jmir.org/2026/1/e94518
doi: 10.2196/94518

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 12.Mar.2026