ACS Partner Journal

ENVIRONMENT & HEALTH

Open Access

pubs.acs.org/EnvHealth

Viewpoint

Ethical Frameworks for Data-Driven Environmental Health Studies in the AI Era

Published as part of Environment & Health special issue "Artificial Intelligence and Machine Learning for Environmental Health".

Miao Yu,* Mingliang Fang, and Bin Wang

Cite This: *Environ. Health* 2025, 3, 443−445

Read Online

ACCESS | Metrics & More | Article Recommendations

The rapid advancement of environmental sensing technologies and artificial intelligence (AI) has ushered in a new era of data-driven environmental health research, especially for the rapid development of exposomics.[1,2] This surge in data collection and analysis capabilities brings unprecedented opportunities for scientific discovery, but also raises critical ethical concerns. Data ethics, the moral framework guiding data management, has become crucial for environmental researchers. The proliferation of advanced instruments, low-cost sensors, and digitalized knowledge has led to an explosion of environmental data. Concurrently, AI models can now derive complex patterns from these vast data sets without traditional hypothesis testing and features extraction, revolutionizing investigations into environmental health issues. However, these advancements bring challenges. Regulations like the EU's General Data Protection Regulation (GDPR) have set new standards for data protection,[3] highlighting the need for robust ethical frameworks in environmental health research. This study aims to explore key ethical considerations in data-driven environmental health studies, focusing on three main areas: data collection, analysis, and sharing. We propose a checklist of ethical guidelines for researchers, building upon existing frameworks.[4] By addressing these ethical challenges, we can promote responsible data practices that maximize the benefits of AI and big data while maintaining scientific integrity and protecting individual privacy.

CHECKLIST

- Researchers should get updated training on data ethics
- Institutional Review Boards (IRB) approval is required for human related studies
- Informed consent is required for human source data/samples
- Sensitive data should be licesened for reuse
- Perform reproducible research
- Evaluate foundation model to avoid transfer learning bias
- Explainable AI should be implemented to enhance transparency and accountability
- Follow FAIR (Findable, Accessible, Interoperable, Reusable) principle to share data
- Data should be deposed in secured repository with encryption

DATA COLLECTION

Environmental health studies typically involve both observational and experimental data. For human-related data, environmental epidemiology researchers have established ethical guidelines,[5] including ensuring no harm to individuals, obtaining informed consent before research initiation, and submitting protocols for review by IRB. Researchers should meticulously record and archive data collection details, such as sampling information, experimental design, questionnaires, and consent statements, for tracking purposes. These protocols should be published alongside research articles or referenced to established methodologies.

Prior to data collection, it is crucial to clarify the intellectual property rights and licenses for the data. For AI training purposes, ensuring the data set represents the target population is vital to prevent biases in model outcomes. As research increasingly relies on portable devices or passive data collection methods, the intended research usage must be clearly and directly addressed in informed consent procedures.

Citizen science data, while valuable, requires careful evaluation of accuracy, consent, and representation before analysis. Simulated, resampled, or augmented data derived from small sample sizes should be clearly labeled, especially for AI-generated or modified content. For Internet-based environmental health studies, researchers should adhere to established ethical guidelines, such as those outlined in the *Internet Research: Ethical Guidelines*.[6]

DATA ANALYSIS

Data analysis in environmental health studies encompasses data preprocessing, statistical analysis, and model training,

Received: December 22, 2024
Accepted: January 27, 2025
Published: February 6, 2025

https://doi.org/10.1021/envhealth.4c00273
Environ. Health 2025, 3, 443−445

validation, and prediction. Each stage presents unique ethical challenges that researchers must address.

In the preprocessing stage, protecting personal information is paramount. Researchers should remove or encrypt personal identifiers, and if such information is essential for the study, prepare a secure codebook for controlled access. To mitigate bias, consider using simulated data or downsampling techniques to balance underrepresented subgroups.

The choice of software or programming language can inadvertently introduce bias. For instance, different statistical packages may use varying default settings for analyses like ANOVA (R implement Type I, python use Type II, and SAS choose Type III) to calculate the sums of squares.[7] To ensure transparency and reproducibility, researchers should document and share software information, version numbers, and analysis scripts. Open-source software is preferable to avoid paywalls that might hinder validation by other researchers.

When using transfer learning or building models on top of existing ones, it is crucial to assess the risk of bias propagation from the foundation models.[8] Researchers should document and release model architecture, algorithms, and hyperparameter optimization processes to facilitate bias tracking.

AI has the potential for misuse. It can expedite the identification of emerging compounds that pose a threat to public health. This same methodology could also be employed to drug discovery which might be linked to potential biochemical weapons.[9,10] As access to powerful AI becomes more common, there's a risk that such research outputs could be misused. In such cases, those kinds of results should be evaluated by individuals from diverse research backgrounds in order to fully mitigate risks before the results are made public.

To address the "black box" nature of some AI models, researchers should implement explainable AI (XAI) techniques.[11] Methods such as Grad-CAM for CNN models or attention visualization for transformer models can help interpret complex models. Perturbation-based methods can also evaluate model performance by modifying inputs, helping to validate models against known ground truths and extract knowledge from the models.[12]

Computational resource usage is an emerging ethical concern. Researchers should optimize code for efficient power utilization and consider the carbon footprint of their analyses. It is also important to credit software developers and prepare code for various computational environments, including cloud and high-performance computing.

DATA SHARING

In recent years, environmental health-related data has increasingly been recognized as a public good,[13] with scientific journals, researchers, and funding agencies advocating for open data sharing. The FAIR principles have emerged as guidelines for scientific data management.[14] However, data sharing raises ethical concerns regarding data ownership, intellectual property, credit attribution, privacy protection, and data security.

Deidentification of shared data is crucial, especially for human samples. Data sharing policies should undergo IRB review and comply with local laws. For instance, researchers in the United States must adhere to the Health Insurance Portability and Accountability Act (HIPAA) when handling protected health information (PHI). However, even with identifiers removed, there's a risk of reidentification, particularly with omics data.[15] To address this, researchers should consider implementing advanced protection measures such as homomorphic encryption, privacy-preserving computation, and permission-based access controls.

Clear licensing and ownership declarations are essential for shared data. Data should be published in open-source formats and made available for data mining. Researchers are encouraged to publish their data in peer-reviewed data journals, such as Scientific Data, which provide detailed descriptions of data sets. Assigning unique identifiers (e.g., DOIs) to data sets, separate from associated papers, is recommended. For data linked to journal articles, researchers should practice reproducible research, sharing their analysis process through tools like Jupyter Notebook or Quarto.

Data storage and accessibility are critical considerations. Shared data should be deposited in professional, discipline-specific repositories (e.g., Metabolomics Workbench for metabolomics data) or general scientific repositories like Zenodo or FigShare. For sensitive data requiring secure transfer, research cyberinfrastructures like Globus can provide HIPAA-compliant solutions.

As AI-generated or modified data becomes more prevalent, implementing mechanisms to detect and label such data may become necessary. This ensures transparency and allows users to account for potential biases or limitations in AI-processed data.

As we've discussed, ethical considerations permeate every stage of the research process, from data collection and analysis to sharing and storage. The responsible management of research data extends far beyond text, figures, or audio; it encompasses vast data sets containing hidden patterns that require careful handling and interpretation. By adhering to ethical frameworks and guidelines, researchers can harness the power of AI and big data while mitigating potential risks. These include preventing biased health condition prediction models, ensuring equitable treatment and policy recommendations, and safeguarding personal information from unintended disclosure.

The checklist we've proposed serves as a starting point for researchers to integrate ethical considerations into their work. However, it is crucial to recognize that data ethics is an evolving field. As technology advances and new ethical challenges emerge, researchers must remain vigilant and adaptable. Moreover, the implementation of ethical data practices is not just a matter of compliance; it is fundamental to the integrity and credibility of environmental health research. By prioritizing transparency, fairness, and privacy protection, researchers can build trust with study participants, fellow scientists, and the public.

In conclusion, we urge environmental health researchers to embrace these ethical frameworks as integral components of their research methodology. By doing so, we can ensure that the powerful tools of data science and AI serve to advance our understanding of environmental health while upholding the highest standards of scientific and ethical conduct.

AUTHOR INFORMATION

Corresponding Author

Miao Yu − *The Jackson Laboratory, Farmington, Connecticut 06032, United States;* orcid.org/0000-0002-2804-6014; Email: miao.yu@jax.org

https://doi.org/10.1021/envhealth.4c00273
Environ. Health 2025, 3, 443−445

Authors

Mingliang Fang − *Department of Environmental Science and Engineering, Fudan University, Shanghai 200433, China*

Bin Wang − *Institute of Reproductive and Child Health/ National Health Commission Key Laboratory of Reproductive Health, School of Public Health, Peking University, Beijing 100191, China; Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing 100191, China;* orcid.org/0000-0002-1164-8430

Complete contact information is available at: https://pubs.acs.org/10.1021/envhealth.4c00273

Notes

The authors declare no competing financial interest.

Biography

Miao Yu, Ph.D., is an Associate Computational Scientist at The Jackson Laboratory. He holds a Ph.D. in Environmental Sciences from the Chinese Academy of Sciences and completed postdoctoral training at the University of Waterloo and the Icahn School of Medicine at Mount Sinai. With a distinctive interdisciplinary background that integrates wet lab experimentation and computational research. Dr. Yu pioneered the concept of "reactomics," a novel approach to uncovering general chemical relationships among molecules in biological samples. He has developed computational workflows, including the "gatekeeper" workflow for identifying exposure-sensitive metabolites and an active molecular network discovery workflow for exposomics studies, integrating advanced machine learning techniques. Dr. Yu is also the author and maintainer of several CRAN/Bioconductor R packages and a reproducible Docker image, all designed for high-resolution mass spectrometry data analysis. His work bridges computational and experimental sciences, driving advancements in data-driven environmental health research.

REFERENCES

(1) Fang, M.; Hu, L.; Chen, D.; Guo, Y.; Liu, J.; Lan, C.; Gong, J.; Wang, B. Exposome in Human Health: Utopia or Wonderland? *The. Innovation* **2021**, *2* (4), 100172.

(2) Zhao, F.; Li, L.; Lin, P.; Chen, Y.; Xing, S.; Du, H.; Wang, Z.; Yang, J.; Huan, T.; Long, C.; Zhang, L.; Wang, B.; Fang, M. HExpPredict: In Vivo Exposure Prediction of Human Blood Exposome Using a Random Forest Model and Its Application in Chemical Risk Prioritization. *Environ. Health Perspect.* **2023**, *131* (3), No. 037009.

(3) General Data Protection Regulation (GDPR) − Legal Text. *General Data Protection Regulation (GDPR)*. https://gdpr-info.eu/ (accessed 2024−12−16).

(4) Zhu, J.-J.; Boehm, A. B.; Ren, Z. J. Environmental Machine Learning, Baseline Reporting, and Comprehensive Evaluation: The EMBRACE Checklist. *Environ. Sci. Technol.* **2024**, *58* (45), 19909−19912.

(5) Kramer, S.; Soskolne, C. L.; Mustapha, B. A.; Al-Delaimy, W. K. Revised Ethics Guidelines for Environmental Epidemiologists. *Environ. Health Perspect.* **2012**, *120* (8), a299−a301.

(6) AoIR Staff. *AoIR's Internet Research Ethics 3.0*. https://aoir.org/ire30/ (accessed 2024−12−16).

(7) Herr, D. G. On the History of ANOVA in Unbalanced, Factorial Designs: The First 30 Years. *Am. Stat.* **1986**, *40* (4), 265−270.

(8) Salman, H.; Jain, S.; Ilyas, A.; Engstrom, L.; Wong, E.; Madry, A. When Does Bias Transfer in Transfer Learning? *arXiv*, July 6, 2022. DOI: 10.48550/arXiv.2207.02842.

(9) Skinnider, M. A.; Wang, F.; Pasin, D.; Greiner, R.; Foster, L. J.; Dalsgaard, P. W.; Wishart, D. S. A Deep Generative Model Enables Automated Structure Elucidation of Novel Psychoactive Substances. *Nat. Mach. Intell.* **2021**, *3* (11), 973−984.

(10) Urbina, F.; Lentzos, F.; Invernizzi, C.; Ekins, S. Dual Use of Artificial-Intelligence-Powered Drug Discovery. *Nat. Mach. Intell.* **2022**, *4* (3), 189−191.

(11) Li, X.; Xiong, H.; Li, X.; Wu, X.; Zhang, X.; Liu, J.; Bian, J.; Dou, D. Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond. *arXiv*, July 15, 2022. DOI: 10.48550/arXiv.2103.10689.

(12) Ivanovs, M.; Kadikis, R.; Ozols, K. Perturbation-Based Methods for Explaining Deep Neural Networks: A Survey. *Pattern Recognit. Lett.* **2021**, *150*, 228−234.

(13) Sabatello, M.; Martschenko, D. O.; Cho, M. K.; Brothers, K. B. Data Sharing and Community-Engaged Research. *Science* **2022**, *378* (6616), 141−143.

(14) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 'tHoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3* (1), No. 160018.

(15) Shringarpure, S. S.; Bustamante, C. D. Privacy Risks from Genomic Data-Sharing Beacons. *Am. J. Hum. Genet.* **2015**, *97* (5), 631−646.