ScienceBank

DOI: 10.61340/FBDTPM
Submitted: 8 May 2025
Accepted after peer review: 10 July 2025
Published: 22 July 2025

✦ Editorial

From big data to personalized medicine: bioinformatics perspectives and challenges

Ilya Ioshikhes

Former professor at Humber College
Correspondence: iioschik@uottawa.ca

Keywords: big data, omics, personalized medicine, formal fallacies, statistics, Pareto principle

Abstract

The rapid accumulation of various types of -omics data is opening new vistas for the development of novel therapeutic approaches grounded in big data. Particularly, personalized medicine stands out as one of the most promising goals. However, it faces significant logical and mathematical challenges. These include the ecological fallacy, which refers to making individual predictions based on average population data, and the Pareto principle, which suggests that approximately 80% of outcomes are driven by 20% of inputs. These raise the possibility of a dramatic discrepancy between our expectations for big data–based personalized medicine and its actual effectiveness. What are the practical issues our community must address to bridge the gap between the promise of big data and the reality of personalized medicine?

Our era may be scientifically defined as the era of super-massive data analysis, in which incredibly large amounts of data from various fields of knowledge are examined, leading to profound conclusions through statistical inference.

Generally speaking, statistical estimates are more reliable when they are supported by large amounts of data, in accordance with the Law of Large Numbers.[1] For instance, in medical diagnostics, the use of specific disease symptoms and biochemical test results yields more reliable outcomes when based on a larger number of observations. Of course, it is not only the quantity of data that matters but also its quality, which may be influenced by factors such as data curation, processing, diversity, reproducibility, explainability, selection bias, noisy labels in image banks, genomic linkage errors, etc.

Thus, while the following perspective emphasizes the quantity of data, it also acknowledges quality as an indispensable component.

Billions of individuals could potentially improve their health through tailored treatments.[2] People would gain access to individualized medicine,[3,4] with care plans designed specifically for them based on their unique genotype, other -omics data, and insights derived from statistical analyses of next-generation large-scale datasets. This approach would not only aid in curing diseases but also in preventing them in a manner best suited to each individual patient.

But are statistical results truly applicable to individuals? To begin exploring this question, let us first consider how it applies in other areas of science.

Statistical physics examines the distribution of parameters within a statistical ensemble (e.g., the energies of gas molecules). Although these parameters vary over time and from one molecule to another, the overall statistical distribution remains stable, resulting in consistent average values such as energy and temperature for the ensemble. Clearly, this is not the case for each individual molecule at any given moment; the behavior of average statistical parameters differs from that of the corresponding parameters at the level of individual molecules.

Eukaryotic gene promoters contain a variety of DNA promoter elements, some more universal than others. Historically, the TATA box was the first general promoter element to be discovered, and all initially described eukaryotic promoters contained it. Later, TATA-less promoters were identified, leading to the hypothesis that another general promoter element must be present in these sequences. Such an element, known as the Initiator (Inr), was soon discovered. However, as more promoter sequences were collected, it became evident that many promoters lack both the TATA box (in over 80% of cases) and the Inr. Although hundreds of additional promoter elements have since been identified, none are, on their own, necessary or sufficient for a genomic sequence to function as a promoter. Still, TATA box and Inr are the most important promoter elements from a statistical perspective. Many individual promoters don't contain them, and other promoter elements are even less universal. This highlights a clear discrepancy between the general functioning of the gene transcription machinery and its operation at the level of individual promoters.

More systematically, similar phenomena have been defined as the ecological (inference) fallacy.[5, 6] This is a logical error that arises when conclusions about individual members of a group are drawn from aggregate data about the group as a whole. The fallacy lies in assuming that what holds true for a population must also apply to each of its individual members, an assumption that can lead to inaccurate or misleading conclusions.

In the case of omics big data, the number of parameters associated with each individual is astronomical (and includes all nucleotides in their genome, the characteristics of their single cells, environmental conditions, and more). This number is much greater when considering the entirety of humankind, or other large communities such as soil bacterial populations. As a result, even with the most powerful computers, analyzing omics big data requires selecting only the most statistically significant features, usually achieved through dimensionality reduction.[7] This involves omitting the majority of features from further analysis, based on the assumption that the remaining parameters still capture the most important features of the system, while making computational analysis cheaper and more feasible.[8,9]

Based on the key parameters selected in the previous step and the results of computational analysis for a given population (defined socially, geographically, etc.), each individual within that population may be assigned certain scores. These scores can guide the selection of medical treatments tailored to the individual's unique characteristics. Such treatments would be (almost) perfectly suited to the person, informed by the vast amounts of relevant big data. We would require data on patient responses to therapies to achieve this level of stratification and personalized treatment. This approach paves the way for the development of tailored therapeutic strategies, an approach known as individualized medicine.

The promise of individualized medicine should benefit all of us, but unfortunately, this is precisely where the ecological fallacy comes into play: "inferring about the nature of an entity (individual -II) based solely upon aggregate statistics collected for the group (population -II) to which that entity belongs".[10] The ecological fallacy is a type of formal fallacy that may arise in the interpretation of statistical data, as discussed above. A formal fallacy, in turn, is a line of reasoning invalidated by a flaw in its logical structure.[11] So, where exactly was the flaw in the reasoning outlined in the previous paragraph? According to Freedman, "the ecological fallacy consists in thinking that relationships observed for groups necessarily hold for individuals".[12] Rather, parameter correlations for the population and for the individual are not necessarily the same, and perhaps are, in fact, usually not the same.

To understand how significant this difference can be, let us consider the Pareto Principle, also known as the 80/20 rule. Originally, economist Vilfredo Pareto observed that 80 % of the land in the Kingdom of Italy was owned by 20 % of the population at that time.[13] This observation was later generalized into what is now known as the Pareto Principle, that in many cases, 80 % of the effects result from 20 % of the causes.[14] This principle has been validated across various fields, including business, economics, and quality control, e.g., the finding that 80 % of meaningful outcomes often come from just 20% of a company's employees.

Although this principle has not yet been widely confirmed in omics data, it has been demonstrated in at least some contexts.[15-19] If applied to our case, it could suggest that 80 % of medical recommendations are derived from just 20% of the population. But what about the remaining 80 %? For them, these recommendations likely do not work so well. Otherwise, they would fall within that initial 20 %. As a result, only about 20 % of new patients might reasonably expect that big data-based individualized medicine will be effective for them. Clearly, this falls short of the efficiency gains we aspire to achieve through big data analysis.

Thus, rather than relying solely on statistics, it might be imperative to use more profound machine learning (ML) and artificial intelligence (AI)-driven models that can capture individual patient traits and drive personalized predictions. These models can better stratify patient populations and perform cross-validations that improve generalizability to the 80 % portion of the population and help detect unseen trends. However, a fundamental problem remains unresolved: while training ML and AI models on big data may enhance diagnosis and treatment at the population level, will it truly improve outcomes for each individual?

Furthermore, while the ecological fallacy and the Pareto principle illustrate fundamental errors in inference and the distribution of effects, they represent only a fraction of the obstacles that separate big data from personalized medicine (PM). In practice, the following technical issues must also be addressed (this list is not exhaustive):

1. **Data quality and curation:** Data curation, or data cleaning, is a basis of the data quality assessment and a key prerequisite of any data analytics services. Given the massive volumes of data to be analyzed, medicine would benefit greatly from automated data curation systems. The analysis of heterogeneous, multidisciplinary data requires the "detection and tracking of inconsistencies, missing values, outliers, and similarities," as well as data standardization processes.[20]
2. **Lack of collection standards, inconsistent labels, missing variables, and selection bias in electronic health records (EHRs) and biobanks:** Since medical decisions are made based on heterogeneous information and criteria provided by patients, providers, and health systems, applying standard missing data techniques is often problematic.[21] Standardized data provenance (the documentation of data origins, and its generation and processing history) based on blockchain technology shows promise. However, several challenges still remain, such as provenance security, communication overhead, scalability, revocation, and General Data Protection Regulation compliance.[22]
3. **Need for robust cleaning, harmonization, and continuous audit pipelines:** The continuous growth of digital datasets originating from diverse sources (sensor networks, social media, business transactions, etc.) demands the development of effective data management strategies. These must include robust data cleansing pipelines to detect and repair dirty data and to address inherent uncertainties caused by noise, missing values, inconsistencies, and other issues.[23]
4. **Representativeness and equity:** Many groups that are underrepresented or excluded in clinical research due to gender, age, race, or ethnicity have been shown to exhibit distinct disease presentations and responses to drugs or therapies. These differences may result from specific genotypes, epigenetic effects, or non-genetic lived experiences such as unconscious racism and variable socioeconomic and educational levels. Proper representativeness is critical for generalizability and population-specific relevance of medical findings.[24]
5. **Underrepresentation of ethnic minorities:** Research has shown that individuals from ethnic minority groups participate in medical research at lower rates than their white counterparts. This may result in invalid treatment generalization across all ethnic groups, without understanding how minority populations are affected, and may contribute to health inequalities and maltreatment for these populations.[25]
6. **Data balancing strategies:** Big data analytics in healthcare involves large-scale data from EHRs, wearable devices, and genomic studies, and can improve clinical decision-making, manage disease outbreak, and personalize treatment plans. However, it also introduces significant security and privacy challenges, with sensitive patient information becoming a target for cyber threats, data breaches, and unauthorized access. Hence, the benefits of data-driven medicine should be balanced with stringent data security, regulatory compliance, and ethical considerations.[26]
7. **Multidimensional integration:** The numerous benefits of big data analysis in healthcare settings are inhibited by multiple technological and cultural obstacles, privacy issues, and the critical need for standardized procedures to efficiently explore the huge volume of healthcare data[27] coming from databases, EHRs and other documents, medical devices and sensors.[28] Dimensionality reduction is indispensable for such analysis,[29, 30] yet may result in loss of crucial individual information necessary for PM.
8. **Challenge of combining genomics, proteomics, clinical and lifestyle data in real time:** While the generation and integration of multi-omics data facilitate PM, the complexity of data integration

across different omics layers, the need for advanced computational tools, the high cost of comprehensive data generation, and issues related to data privacy, standardization, and the need for robust validation in diverse populations remain significant obstacles. Real-time integration of genomic data into clinical workflows will allow clinicians to use this data for personalized diagnosis, treatment planning, and preventive care.[31]

9. **Interoperability and data governance:** To transform medical decision-making, an unprecedented amount of data must be collected across diverse sectors. Standards for data sharing and interoperability, along with robust data stewardship and governance models, should be established to ensure the protection and integrity of the public health data system.[32]
10. **Overfitting in datasets:** ML models may outperform statistical regression algorithms in clinical predictions. Proper validation and careful tuning of these models are essential to avoid overfitting, where the model closely resembles the training dataset, and to ensure strong generalizability.[33]
11. **Lack of explicit causality:** In medicine, causality is generally expressed probabilistically, that is, a specific event carries a certain probability of leading to a specific outcome, as determined through randomized controlled trials (RCTs). For example, exposure to a risk factor (like cigarette smoking or diesel emissions) increases the probability of developing lung cancer, whereas certain treatment (such as a specific type of chemotherapy) increases the likelihood of survival in patients with a cancer diagnosis, although does not guarantee it. If RCTs are not feasible for practical or ethical reasons, observational studies should be used.[34]
12. **Limited explainability:** Although ML-based AI applications in healthcare can provide numerous benefits, they also present significant issues concerning explainability. These models often resemble a "black box," offering no clear reasoning for a given conclusion. This makes understanding and justifying the models' outcomes extremely difficult and adds an extra level of complexity to the decision-making process.[35]
13. **Security, privacy, and regulation:** The digitization of healthcare brings numerous benefits, including improved efficiency, enhanced patient care, and advanced data analytics. However, it also raises significant concerns regarding security and privacy, requiring the appropriate collection, use, and sharing of personal health information while preserving patient's autonomy and confidentiality. The latter require robust authentication and access controls, encryption and data protection measures, continuous security monitoring, privacy-by-design principles, and effective training and awareness programs[36]
14. **Hospital infrastructure barriers, staff training:** The complexity of PM approaches and their associated challenges requires the development of effective engagement strategies, collaboration frameworks, infrastructure, education and training programs, ethical policies, resource allocation guidelines, regulatory compliance measures, and robust data management and privacy practices. All of these must be implemented by appropriately trained staff.[37]

Our growing community is focused on exploring the challenges and opportunities that arise at the intersection of big data and PM. As this field continues to evolve, contributions from statistics, data analysis, and biomedical research play a key role in shaping its direction. As we confront these intellectually demanding and practically impactful challenges, what perspectives (disciplinary, methodological, or philosophical) will help move the field forward?

Acknowledgements

The author wants to thank Boris Dashevsky for helpful discussions.

Conflicts of interest

This editorial was commissioned by ScienceBank with the author receiving compensation for their contribution. Topic selection and development remained entirely at the author's discretion, and the manuscript underwent an independent peer review process in accordance with established editorial policies.

Funding

This editorial was commissioned by ScienceBank with the author receiving compensation for their contribution. Topic selection and development remained entirely at the author's discretion, and the manuscript underwent an independent peer review process in accordance with established editorial policies.

References

1. Michel Dekking (2005). *A Modern Introduction to Probability and Statistics*, 181–90. Springer. DOI: 10.1007/1-84628-168-7

2. Joel T. Dudley, Konrad J. Karczewski (2014). *Exploring Personal Genomics.* Oxford University Press. DOI: 10.1093/acprof:oso/9780199644483.001.0001

3. Melinda L. Telli, Sharon A. Hunt, Robert W. Carlson, et al. (2007). "Trastuzumab-related cardiotoxicity: calling into question the concept of reversibility." *Journal of Clinical Oncology* 25: 3525–33. DOI: 10.1200/JCO.2007.11.0106

4. Giuseppe Saglio, Alessandro Morotti, Giovanna Mattioli, et al. (2004). "Rational approaches to the design of therapeutics targeting molecular markers: the case of chronic myelogenous leukemia." *Annals of the New York Academy of Sciences* 1028: 423–31. DOI: 10.1196/annals.1322.050

5. Charles Ess, Fay Sudweeks (2001). *Culture, Technology, Communication: Towards an Intercultural Global Village*, 90. SUNY Press. DOI: 10.1515/9780791490488

6. Gary King (1997). *A Solution to the Ecological Inference Problem.* Princeton University Press.

7. Laurence van der Maaten, Eric Postma, Jaap van den Herik (Oct 26, 2009). "Dimensionality Reduction: A Comparative Review." *Journal of Machine Learning Research* 10: 66–71.

8. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013). *An Introduction to Statistical Learning*, 204. Springer. DOI: 10.1007/978-1-4614-7138-7

9. Janez Brank, Dunja Mladenić, Marko Grobelnik, et al. (2011). "Feature Selection." In Claude Sammut and Geoffrey I. Webb (eds.), *Encyclopedia of Machine Learning*, 402–6. Springer US.

10. David H. Fischer (1970). *Historians' Fallacies: Toward a Logic of Historical Thought.* Harper & Row.

11. Stephen F. Barker (1989). “Chapter 6: Fallacies.” In *The Elements of Logic.* McGraw-Hill.

12. David A. Freedman (1999). “Ecological Inference and the Ecological Fallacy.” *International Encyclopedia of the Social & Behavioral Sciences*, Technical Report No. 549. https://web.stanford.edu/class/ed260/freedman549.pdf

13. Vilfredo Pareto (1896–1897). *Cours d’Économie Politique*. F. Rouge and F. Pichon.

14. Juran (2019). *A Guide to the Pareto Principle (80/20 Rule & Pareto Analysis). Juran Blog.* https://www.juran.com/blog/a-guide-to-the-pareto-principle-80-20-rule-pareto-analysis/

15. Xijiang Yu, Theo H. E. Meuwissen (2011). “Using the Pareto Principle in Genome-Wide Breeding Value Estimation.” *Genetics Selection Evolution* 43: 35. DOI: 10.1186/1297-9686-43-35

16. Cédric Saule, Robert Giegerich (2015). “Pareto Optimization in Algebraic Dynamic Programming.” *Algorithms for Molecular Biology* 10: 22. DOI: 10.1186/s13015-015-0051-7

17. Yuping Li, Dmitri A. Petrov, Gavin Sherlock (2019). “Single Nucleotide Mapping of Trait Space Reveals Pareto Fronts that Constrain Adaptation.” *Nature Ecology & Evolution* 3: 1539–51. DOI: 10.1038/s41559-019-0993-0

18. Linsong Dong, Ming Fang, Zhiyong Wang (2017). “Prediction of Genomic Breeding Values Using New Computing Strategies for the Implementation of MixP.” *Scientific Reports* 7: 17200. DOI: 10.1038/s41598-017-17366-2

19. Irene Otero-Muras, Julio R. Banga (2016). “Exploring Design Principles of Gene Regulatory Networks via Pareto Optimality.” *IFAC-PapersOnLine* 49: 809–14. DOI: 10.1016/j.ifacol.2016.07.289

20. Vasileios C. Pezoulas, Konstantina D. Kourou, Fanis Kalatzis, et al. (2019). “Medical Data Quality Assessment: On the Development of an Automated Framework for Medical Data Curation.” *Computers in Biology and Medicine* 107: 270–83. DOI: 10.1016/j.compbiomed.2019.03.00

21. Sarah B. Peskoe, David Arterburn, Karen J. Coleman, et al. (2021). “Adjusting for Selection Bias Due to Missing Data in Electronic Health Records-Based Research.” *Statistical Methods in Medical Research* 30: 2221–38. DOI: 10.1177/09622802211027601

22. Mansoor Ahmed, Amil R. Dar, Marcus Helfert, et al. (2023). “Data Provenance in Healthcare: Approaches, Challenges, and Future Directions.” *Sensors* 23: 649. DOI: 10.3390/s23146495

23. Mehdi Hosseinzadeh, Elham Azhir, Omed Hassan Ahmed, et al. (2023). “Data Cleansing Mechanisms and Approaches for Big Data Analytics: A Systematic Study.” Journal of Ambient Intelligence and Humanized Computing 14: 99–111. DOI: 10.1007/s12652-021-03590-2

24. National Academies of Sciences, Engineering, and Medicine (2022). *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups.* National Academies Press.

25. Shahina Pardhan, Tarjit Sehmbi, Rumalie Wijewickrama, et al. (2025). “Barriers and Facilitators for Engaging Underrepresented Ethnic Minority Populations in Healthcare Research: An Umbrella Review.” *International Journal for Equity in Health* 24: 70. DOI: 10.1186/s12939-025-02431-4

26. Danish Khan, Madison Daena (2025). “Balancing Innovation and Security: Navigating the Challenges of Big Data Analytics in Healthcare IT.”

27. João Pedro Cajado (2025). “Exploring the Benefits and Challenges of Integrating Big Data and Data Analysis in Healthcare: A Systematic Literature Review.” In Pereira, R., Bianchi, I., Rocha, A.

(eds.), *Digital Technologies and Transformation in Business, Industry and Organizations. Studies in Systems, Decision and Control*, vol. 577. Springer. DOI: 10.1007/978-3-031-78412-5_6

28. Kornelia Batko, Andrzej Ślęzak (2022). "The Use of Big Data Analytics in Healthcare." *Journal of Big Data* 9: 3. DOI: 10.1186/s40537-021-00553-4

29. Joshua T. Vogelstein, Eric W. Bridgeford, Minh Tang, et al. (2021). "Supervised Dimensionality Reduction for Big Data." *Nature Communications* 12: 2872. DOI: 10.1038/s41467-021-23102-2

30. Diaa S. Abdelminaam, Ahmed A. Alheijat, Mohamed Taha (2024). "A Comprehensive Survey on Dimension Reduction for Medical Big Data Using Optimization Algorithms." *2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, Cairo, Egypt, 331–37. DOI: 10.1109/MIUCC62295.2024.10783491

31. Getnet Molla, Molalegne Bitew (2024). "Revolutionizing Personalized Medicine: Synergy with Multi-Omics Data Generation, Main Hurdles, and Future Perspectives." *Biomedicines* 12: 2750. DOI: 10.3390/biomedicines12122750

32. Laurie T. Martin, Christopher Nelson, Douglas Yeung, et al. (2022). "The Issues of Interoperability and Data Connectedness for Public Health." *Big Data* 10(Suppl 1): S19–S24. DOI: 10.1089/big.2022.0207

33. Paris Charilaou, Robert Battat (2022). "Machine Learning Models and Over-Fitting Considerations." *World Journal of Gastroenterology* 28: 605–7. DOI: 10.3748/wjg.v28.i5.605

34. Emilio A. L. Gianicolo, Martin Eichler, Oliver Muensterer, et al. (2020). "Methods for Evaluating Causality in Observational Studies." *Deutsches Ärzteblatt International* 117: 101–7. DOI: 10.3238/arztebl.2020.0101

35. Claudia Giorgetti, Giuseppe Contissa, Giuseppe Basile (2025). "Healthcare AI, Explainability, and the Human-Machine Relationship: A (Not So) Novel Practical Challenge." *Frontiers in Medicine (Lausanne)* 12: 1545409. DOI: 10.3389/fmed.2025.1545409

36. Lubna Abdel Jawad (2024). "Security and Privacy in Digital Healthcare Systems: Challenges and Mitigation Strategies." *Abhigyan* 42: 23–31. DOI: 10.1177/09702385241233073

37. Caterina Cinti, Maria Giovanna Trivella, Michael Joulie, et al. (2024). "The Roadmap toward Personalized Medicine: Challenges and Opportunities." *Journal of Personalized Medicine* 14: 546. DOI: 10.3390/jpm14060546