Data Stewardship and Curation Practices in AI-Driven Genomics and Automated Microscopy Image Analysis for High-Throughput Screening Studies: Promoting Robust and Ethical AI Applications

Anonymous Author(s)
Affiliation
Address
email

Abstract

The increasing adoption of AI and next-generation sequencing (NGS) has revolutionized genomics and high-throughput screening (HTS), transforming how cellular processes and disease mechanisms are understood. However, these advancements generate vast datasets requiring effective data stewardship and curation practices to maintain data integrity, privacy, and accessibility. This review consolidates existing knowledge on key aspects, including data governance, quality management, privacy measures, ownership, access control, accountability, traceability, curation frameworks, and storage systems. Major challenges such as managing biases, ensuring data quality, and securing privacy are highlighted. Advanced cryptographic techniques, federated learning, and blockchain technology are proposed as strategic solutions, emphasizing standards compliance, ethical oversight, and tailored access control frameworks. Effective data stewardship is vital for advancing AI-driven genomics and microscopy research. Stakeholders must prioritize robust data governance and privacy measures to ensure data integrity and ethical use. Collaborative efforts should focus on developing transparent data-sharing policies and interoperable platforms to foster innovation and advance research practices. The study promotes collaboration among researchers, robust data governance, privacy and security, clear policies, and educational initiatives to prepare future researchers.

1 Introduction

Advancements in AI and next-generation sequencing (NGS) have revolutionized genomics and high-throughput screening (HTS) studies, enabling the integration of multi-dimensional data [6, 4, 5]. Automated high-content screening (HCS) methodologies, combining microscopy image acquisition and analysis, are now pivotal for understanding cellular processes and assessing drug efficacy [3, 2]. However, these technologies generate vast datasets that require robust data stewardship and curation practices to ensure their reliability and accessibility. Therefore, this study aimed to elucidate best practices in data stewardship and curation for AI-driven genomics and automated microscopy image analysis within high-throughput screening studies.

2 Methods

A systematic literature search was conducted up to December 30, 2023, across PubMed, MEDLINE, Embase, Scopus, and Web of Science. The search focused on data governance, curation frameworks,

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

algorithmic bias, and data storage. Realist synthesis methodology was used to integrate diverse theoretical frameworks, with three independent reviewers. The review process included six stages, starting with an extensive search across multiple research databases, resulting in 273 documents. This was followed by screening based on broad criteria, titles, abstracts, and full texts, which narrowed the pool to 38 highly relevant citations.

3 Experimental Results

Our findings revealed a significant surge in research activity in recent years, particularly in 2023, reflecting the increasing recognition of the importance of robust data governance frameworks. Notably, while 36 articles extensively discussed data interoperability and sharing measures, areas such as model explainability and data augmentation remained underexplored, highlighting crucial gaps that need to be addressed. The integration of diverse data types, including sequencing, clinical, proteomic, and imaging data, underscored the complexity and breadth of AI applications in genomics and microscopy.

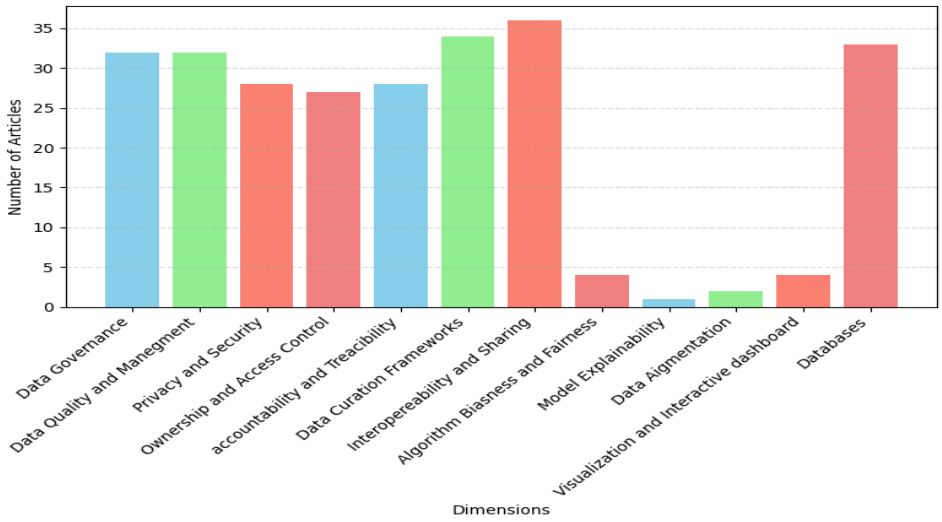

Figure 1: Illustrated the number of articles addressed the data stewardship and curation practice through different dimensions

Moreover, our review emphasized that infrastructure optimization, ethical considerations, access control mechanisms, and transparent data sharing policies were the most critical challenges in AI-based data stewardship. Advanced cryptographic techniques, federated learning, and blockchain technology were proposed to address challenges like data quality, privacy, and bias management. We identified robust data governance measures, such as GA4GH standards[3], DUO versioning, and attribute-based access control[2], as essential for ensuring data integrity, security, and ethical use. The importance of Data Management Plans (DMPs) [1], meticulous metadata curation, and advanced cryptographic techniques emerged as pivotal in mitigating data security and identifiability risks.

4 Conclusion

These findings provide a comprehensive overview of current practices and challenges in data stewardship, offering a roadmap for enhancing the robustness and ethical standards of AI applications in genomics and microscopy. Effective data stewardship and curation are vital for advancing AI-driven genomics and microscopy image analysis. Prioritizing robust governance, quality management, and secure sharing frameworks is essential. Collaborative efforts must focus on developing transparent data-sharing policies and interoperable platforms to foster innovation and advance research practices.

References

[1] Faisal M Fadlelmola, Lyndon Zass, Melek Chaouch, Chaimae Samtal, Verena Ras, Judit Kumuthini, Sumir Panji, and Nicola Mulder. Data management plans in the genomics research revolution of africa: Challenges and recommendations. *Journal of biomedical informatics*, 122:103900, 2021.

[2] David Reddick, Justin Presley, Frank Alex Feltus, and Susmit Shannigrahi. Wip: Aabac-automated attribute based access control for genomics data. In *Proceedings of the 27th ACM on Symposium on Access Control Models and Technologies*, pages 217–222, 2022.

[3] Michael C Schatz, Anthony A Philippakis, Enis Afgan, Eric Banks, Vincent J Carey, Robert J Carroll, Alessandro Culotti, Kyle Ellrott, Jeremy Goecks, Robert L Grossman, et al. Inverting the model of genomics data sharing with the nhgri genomic data science analysis, visualization, and informatics lab-space. *Cell Genomics*, 2(1), 2022.

[4] Marcel P Schilling, Razan El Khaled El Faraj, Joaquín Eduardo Urrutia Gómez, Steffen J Sonnentag, Fei Wang, Britta Nestler, Véronique Orian-Rousseau, Anna A Popova, Pavel A Levkin, and Markus Reischl. Automated high-throughput image processing as part of the screening platform for personalized oncology. *Scientific Reports*, 13(1):5107, 2023.

[5] Rachel H Toczydlowski, Libby Liggins, Michelle R Gaither, Tanner J Anderson, Randi L Barton, Justin T Berg, Sofia G Beskid, Beth Davis, Alonso Delgado, Emily Farrell, et al. Poor data stewardship will hinder global genetic diversity surveillance. *Proceedings of the National Academy of Sciences*, 118(34):e2107934118, 2021.

[6] Galen EB Wright, Pieter GJ Koornhof, Adebowale A Adeyemo, and Nicki Tiffin. Ethical and legal implications of whole genome and whole exome sequencing in african populations. *BMC Medical Ethics*, 14:1–15, 2013.