Transparent Reporting for Healthcare GenAI

Arinbjörn Kolbeinsson
K01
Reykjavík, Iceland
arinbjorn@K01.is

Benedikt Kolbeinsson
K01
Reykjavík, Iceland
benedikt@K01.is

Abstract

Generative AI systems in healthcare offer significant potential but face challenges due to the lack of standardized reporting frameworks. This gap hinders reproducibility, evaluation and regulation, potentially compromising patient safety. Existing tools like Data Cards and Model Cards are inadequate for healthcare's specific needs, such as regulatory compliance and clinical relevance. Drawing on the success of the STROBE checklist in epidemiology, we propose an extension tailored for GenAI in healthcare. This framework enhances transparency, reproducibility and safety, supporting the effective integration of GenAI systems into clinical practice.

1 Introduction

The rapid advancement and widespread adoption of Generative Artificial Intelligence (GenAI) systems in healthcare have brought about remarkable innovations and potential benefits. However, this progress is accompanied by a significant challenge: the lack of standardised reporting processes for these systems. This deficiency creates a multitude of issues, including difficulties in reproducibility, challenges in assessing the reliability and safety of GenAI applications, and obstacles to effective regulation and governance. The absence of a uniform reporting framework hinders transparency, impedes proper evaluation, and potentially compromises patient safety in healthcare settings.

Efforts to address similar challenges in other domains have led to the development of tools such as Data Cards (Pushkarna et al., 2022) and Model Cards (Mitchell et al., 2019). These frameworks provide valuable insights into dataset characteristics and model performance, respectively. Data Cards offer a standardised way to document dataset provenance, composition and intended uses, while Model Cards provide a concise overview of a model's performance characteristics, limitations and ethical considerations. While these tools have proven useful in general AI applications, they often fall short in addressing the unique and complex requirements of healthcare applications, where factors such as patient privacy, clinical relevance and regulatory compliance play crucial roles.

For medical research, the STROBE checklist (Vandenbroucke et al., 2007) has been instrumental in standardizing the reporting of observational studies. This comprehensive checklist has significantly improved the quality, transparency and reproducibility of epidemiological research by providing a clear framework for authors to follow when reporting their studies. The success of STROBE in enhancing the rigor and clarity of observational research, along with its extensions for genetic studies (STREGA) (Little et al., 2009) and molecular epidemiology (STROBE-ME) (Gallo et al., 2012), serves as an inspiration and model for addressing the reporting challenges in GenAI health applications.

In this work, we propose an extension to the STROBE checklist specifically tailored for GenAI health applications. This extension aims to bridge the gap between existing AI reporting frameworks and the specialised needs of the healthcare domain. By adapting the proven structure of STROBE and incorporating elements crucial to GenAI systems, we seek to create a comprehensive reporting

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

standard that addresses the unique challenges posed by these technologies. This proposed framework has the potential to enhance transparency, facilitate proper evaluation, improve reproducibility and ultimately contribute to the safe and effective integration of GenAI systems in healthcare.

2 The GenAI reporting guideline: A STROBE Extension

We adopt a structured approach, examining the necessary reporting requirements on a section-by-section basis. We compared and analysed the guidelines from the original STROBE checklist (Vandenbroucke et al., 2007), model cards (Mitchell et al., 2019) and data cards (Pushkarna et al., 2022). Based on these insights, we compiled the following motivations for our reporting guideline:

Incorporating GenAI in healthcare studies introduces a novel dimension that needs to be explicitly communicated. By indicating the use of GenAI and its role in the title and abstract, researchers make it clear to readers what methods were used and why they are relevant. A well-balanced **abstract** also offers insight into the potential impact of GenAI on healthcare outcomes, providing transparency and aiding quick appraisal of the study's significance.

The **introduction** sets the scientific context for using GenAI in healthcare. It is essential to explain the rationale for selecting GenAI as a tool, whether for data synthesis, simulation, or other purposes, alongside describing the specific healthcare problem being addressed. Stating the objectives and hypotheses at this stage ensures the reader understands how the use of GenAI is expected to contribute to solving the defined healthcare challenges and the specific outcomes the research aims to achieve.

Key **methodological elements**, such as the design of the GenAI system, its architecture and data sources (whether real, synthetic or a combination), must be documented early. This allows other researchers to assess the validity of the approach. Detailed descriptions of data collection processes, including how synthetic datasets are generated and validated, are vital for evaluating the quality and robustness of the findings. Moreover, providing a transparent account of how data privacy and ethics were handled, including anonymisation and compliance with privacy regulations such as GDPR or HIPAA, is critical in the sensitive field of healthcare.

A clear definition of the **GenAI model**, including the architecture, training procedures and hyperparameters, allows replication of the study and evaluation of its methodology. The validation process should also be outlined, emphasizing comparison with real-world data or clinical benchmarks to ensure the model's relevance. Efforts to mitigate bias and ensure fairness, particularly across diverse populations, are crucial for maintaining equity in healthcare outcomes and ensuring that GenAI models do not inadvertently disadvantage underrepresented groups.

Presenting the **results** of the GenAI model using appropriate metrics, such as accuracy, sensitivity and specificity, provides a quantifiable measure of its effectiveness. These results should also include an assessment of the impact on healthcare outcomes, enabling the reader to understand the practical implications of the GenAI system. Any biases detected in the model or synthetic data should be reported, along with the strategies used to address them, to ensure the findings are robust and inclusive. Furthermore, documenting the clinical relevance of the GenAI output offers insight into how it might influence real-world healthcare delivery.

The **discussion** should reflect on key findings, comparing the use of GenAI to traditional methods if relevant, thus helping contextualise its contribution to healthcare. A balanced evaluation of the strengths and limitations, particularly regarding scalability, diversity and biases, ensures that the reader can critically assess the applicability of the findings. Regulatory requirements should also be discussed comprehensively, showcasing the precautions taken to ensure compliance.

The **conclusion** should serve to highlight the overall contributions of the research, emphasizing how GenAI improves transparency, reproducibility and impacts healthcare practice. This section reinforces the value of GenAI and its potential to transform healthcare delivery by making data-driven decisions more accessible and scalable.

To support reproducibility, it is necessary to provide supplementary materials, including details on code and data availability. Ensuring that these materials are accessible allows other researchers to replicate the study's findings. Including a checklist of adherence to the GenAI STROBE guidelines enhances transparency, verifying that all relevant aspects of the study have been thoroughly reported.

3 GenAI for Health Reporting Guideline

Section/Topic	Item No.	Recommendation
Title and abstract	1	(a) Indicate the use of GenAI and the design (e.g., synthetic data generation, simulation) in the title or abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found, including the role of GenAI and its impact on healthcare outcomes
Introduction	2	Explain the scientific background, rationale for using GenAI (e.g., for data synthesis, simulation, digital twins) and the healthcare problem being addressed
	3	State specific objectives, including any prespecified hypotheses and the expected impact of GenAI in addressing the healthcare problem
Methods	4	Present key elements of design early in the paper, such as the GenAI model architecture, data sources (real, synthetic, or combined) and validation approach
Data sources/measurement	5	Describe (if known) the sources and methods of data collection, including details on real-world datasets, synthetic data generation and validation processes
Data privacy and ethics	6	(a) Explain how privacy was ensured, including the use of anonymization or de-identification in the synthetic data (b) Describe the process of ethical approval, informed consent (if applicable) and adherence to privacy laws (e.g., GDPR, HIPAA)
GenAI model	7	(a) Clearly define the GenAI model used, including architecture, training process and hyperparameters (b) Explain how the model was validated, including any comparison against real-world data or clinical benchmarks
Bias and fairness	8	Describe any efforts to address potential sources of bias in the GenAI model and synthetic data, and how fairness was ensured across different populations
Results	9	Present the GenAI model's performance using key metrics (e.g., accuracy, specificity, sensitivity for classification), and describe how the synthetic data or predictions impact healthcare outcomes
	10	Report any bias detected and how it was mitigated, particularly with regards to underrepresented populations in the data
	11	Document the clinical relevance of GenAI results, including how they influenced patient outcomes, diagnosis accuracy or treatment efficiency
Discussion	12	Summarise the key findings and their clinical implications, comparing the use of GenAI to traditional methods
Strengths and limitations	13	(a) Highlight the strengths of using GenAI, such as improved scalability or data diversity (b) Address limitations, including potential biases in the data or generalizability concerns
Regulatory compliance	14	Discuss any regulatory requirements and the measures taken to ensure compliance.
Conclusion	15	Describe the overall contribution of GenAI to healthcare, focusing on transparency, reproducibility and real-world impact
Supplementary material	16	(a) Provide details on code and data availability, ensuring that methods are reproducible (b) Include this checklist of adherence to the GenAI STROBE extension guidelines

4 Discussion

We have proposed a comprehensive reporting guideline for GenAI applications in healthcare, building upon the well-established STROBE checklist (Vandenbroucke et al., 2007). This guideline aims to address the critical need for standardization in reporting GenAI use in medical contexts. By improving transparency, facilitating more rigorous evaluation and supporting safety and regulatory compliance, we believe this framework can significantly contribute to the responsible development and deployment of GenAI technologies in healthcare.

The proposed guideline offers several potential benefits:

- Enhanced Reproducibility: By standardizing reporting practices, researchers and developers can more easily reproduce and validate GenAI models and their results.
- Improved Risk Assessment: Comprehensive reporting enables better identification and mitigation of potential risks associated with GenAI use in healthcare.
- Facilitated Regulatory Compliance: A standardised framework can help align GenAI applications with existing and emerging healthcare regulations.
- Increased Trust: Greater transparency in reporting can foster trust among healthcare providers, patients and regulatory bodies.

However, we acknowledge several limitations to our work. Firstly, this is a *rapidly evolving field*. The fast-paced nature of GenAI research and development in healthcare poses a challenge to creating a long-lasting reporting standard. New techniques, architectures and applications may emerge that require updates to the guideline. To address this, we propose regular reviews and updates to the framework, perhaps on an annual or biennial basis.

Secondly, the use cases are *fragmented* and diverse applications of GenAI in healthcare make it challenging to create a one-size-fits-all reporting standard. Specialised factors critical to specific use cases may be overlooked. For example:

- In medical imaging AI, reporting on data augmentation techniques and image preprocessing steps is crucial.
- For natural language processing models in clinical documentation, details on language model fine-tuning and domain-specific vocabulary are essential.
- In drug discovery AI, reporting on molecular representation methods and binding affinity prediction metrics is vital.

To mitigate this, we suggest exploring domain-specific extensions to the core guideline.

Lastly, applications make use of *Closed-Source Models*. The increasing use of closed-source or proprietary models in GenAI applications presents a significant challenge to transparency (Char et al., 2018). While our guideline encourages maximum disclosure, it may be limited in its ability to ensure full transparency for such models. We recommend that users of closed-source models report as much information as possible within the constraints of their agreements with model providers.

Future work to address these limitations and further improve the guideline could include developing a dynamic, digital version of the guideline that can be more easily updated as the field evolves. This would make it easier to further create a series of specialised sub-checklists for specific GenAI applications in healthcare (e.g., imaging AI, clinical NLP, drug discovery). Another important factor is collaborating with regulatory bodies to align the guideline with emerging AI governance frameworks in healthcare Meskó and Topol (2023). Finally, conducting case studies to evaluate the practical implementation and effectiveness of the guideline in real-world GenAI healthcare projects will help validate the guidelines themselves.

This first iteration of the guideline serves as a foundation for further refinement and expansion. We encourage the broader medical and health AI communities to engage with this framework, provide feedback and contribute to its ongoing development. By fostering a collaborative approach, we can ensure that the guideline remains relevant, comprehensive and effective in promoting responsible GenAI use in healthcare.

The integration of GenAI in healthcare holds immense promise for improving patient care, accelerating medical research and enhancing healthcare delivery (Topol, 2019). However, the rapid

advancement and adoption of these technologies necessitate a structured approach to reporting and evaluation to ensure their safe and effective use.

Our proposed extension of the STROBE checklist for GenAI applications in healthcare represents a significant step towards standardizing reporting practices in this critical domain. By providing a comprehensive framework that addresses the unique challenges of GenAI in medical contexts, we aim to foster transparency, facilitate rigorous evaluation and support regulatory compliance.

As we move forward, the continued refinement and adaptation of this guideline will be crucial. We call upon researchers, developers, healthcare professionals and policymakers to engage with this framework, provide feedback and contribute to its evolution. Through collaborative efforts, we can work towards a future where GenAI technologies are deployed responsibly and effectively in healthcare, ultimately leading to improved patient outcomes and advances in medical knowledge.

5 Conclusion

The journey towards standardised reporting for GenAI in healthcare is ongoing, and this guideline represents an important milestone in that journey. As the field continues to evolve, so too must our approaches to ensuring transparency, safety and efficacy. We look forward to the continued development and implementation of these reporting standards and to the positive impact they will have on the future of healthcare.

References

Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 378(11):981–983.

Gallo, V., Egger, M., McCormack, V., Farmer, P. B., Ioannidis, J. P., Kirsch-Volders, M., Matullo, G., Phillips, D. H., Schoket, B., Stromberg, U., et al. (2012). Strengthening the reporting of observational studies in epidemiology–molecular epidemiology (strobe-me): an extension of the strobe statement. *Mutagenesis*, 27(1):17–29.

Little, J., Higgins, J. P., Ioannidis, J. P., Moher, D., Gagnon, F., Von Elm, E., Khoury, M. J., Cohen, B., Davey-Smith, G., Grimshaw, J., et al. (2009). Strengthening the reporting of genetic association studies (strega)—an extension of the strobe statement. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(7):581–598.

Meskó, B. and Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. ACM.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. (2022). Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56.

Vandenbroucke, J. P., Elm, E. v., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., and Initiative, S. (2007). Strengthening the reporting of observational studies in epidemiology (strobe): explanation and elaboration. *Annals of internal medicine*, 147(8):W–163.