Gameiro *et al.*
Journal of Biomedical Science (2025) 32:14
https://doi.org/10.1186/s12929-024-01106-6

NSTC 國家科學及技術委員會 National Science and Technology Council
The cost of publication in *Journal of Biomedical Science* is borne by the National Science and Technology Council (NSTC), Taiwan.

Journal of Biomedical Science

RESEARCH

Open Access

The Data Artifacts Glossary: a community-based repository for bias on health datasets

Rodrigo R. Gameiro[1,2†], Naira Link Woite[1,2†], Christopher M. Sauer[1,3,4†], Sicheng Hao[5], Chrystinne Oliveira Fernandes[1,2], Anna E. Premo[6,7], Alice Rangel Teixeira[8], Isabelle Resli[9], An-Kwok Ian Wong[5] and Leo Anthony Celi[1,2,10*]

Abstract

Background The deployment of Artificial Intelligence (AI) in healthcare has the potential to transform patient care through improved diagnostics, personalized treatment plans, and more efficient resource management. However, the effectiveness and fairness of AI are critically dependent on the data it learns from. Biased datasets can lead to AI outputs that perpetuate disparities, particularly affecting social minorities and marginalized groups.

Objective This paper introduces the "Data Artifacts Glossary", a dynamic, open-source framework designed to systematically document and update potential biases in healthcare datasets. The aim is to provide a comprehensive tool that enhances the transparency and accuracy of AI applications in healthcare and contributes to understanding and addressing health inequities.

Methods Utilizing a methodology inspired by the Delphi method, a diverse team of experts conducted iterative rounds of discussions and literature reviews. The team synthesized insights to develop a comprehensive list of bias categories and designed the glossary's structure. The Data Artifacts Glossary was piloted using the MIMIC-IV dataset to validate its utility and structure.

Results The Data Artifacts Glossary adopts a collaborative approach modeled on successful open-source projects like Linux and Python. Hosted on GitHub, it utilizes robust version control and collaborative features, allowing stakeholders from diverse backgrounds to contribute. Through a rigorous peer review process managed by community members, the glossary ensures the continual refinement and accuracy of its contents. The implementation of the Data Artifacts Glossary with the MIMIC-IV dataset illustrates its utility. It categorizes biases, and facilitates their identification and understanding.

Conclusion The Data Artifacts Glossary serves as a vital resource for enhancing the integrity of AI applications in healthcare by providing a mechanism to recognize and mitigate dataset biases before they impact AI outputs. It not only aids in avoiding bias in model development but also contributes to understanding and addressing the root causes of health disparities.

Keywords Bias, Health equity, Dataset, Data Artifacts Glossary, Artificial intelligence, Machine learning

†Rodrigo R. Gameiro, Naira Link Woite and Christopher M. Sauer contributed equally to this work.

*Correspondence:
Leo Anthony Celi
lceli@mit.edu
Full list of author information is available at the end of the article

Introduction

Artificial Intelligence (AI) has the potential to revolutionize healthcare by offering sophisticated algorithms capable of diagnosing diseases, crafting personalized treatment plans, aiding clinicians in decision-making processes, and alleviating the administrative burden on healthcare practitioners [1, 2]. This technological advancement has been proposed to not only enhance the efficiency of healthcare delivery but also to shift the focus back to patient-centered care [3]. However, achieving this promise is not without its challenges. Deploying AI in healthcare presents complexities unparalleled in other sectors, primarily due to the intricate nature of medical practice, the historical biases ingrained within it and intrinsic epidemiological challenges working with electronic health record data [4].

Bias in research is defined as a systematic error or tendency that prevents impartial consideration by favoring one answer over others [5]. The issue of clinical bias is a well-known topic in the medical field, underscored by extensive research that explores its manifestations within a societal framework marked by inequity, prejudice, and discrimination [6]. These biases have tangible and detrimental effects on patient care, leading to disparities in diagnosis, treatment, and outcomes across diverse populations [7]. For instance, when compared to white Americans, pain management in black Americans is systematically worse, due to false beliefs about biological differences between these two groups [8]. Moreover, racial and ethnic minority patients are less likely to be screened for diabetic retinopathy, even though they are more likely to have poorer glycemic control [9].

Training AI on clinical data derived from a world rife with such biases risks not merely replicating, but also amplifying and perpetuating them [10, 11]. Notwithstanding, it could even reconfigure new ones that would remain elusive due to the inherently opaque nature of some AI algorithms [12]. Already, evidence of bias in AI spans across a variety of applications, from sex-based disparities in algorithms predicting cardiovascular risks [13], to ethnic disparities in the detection of skin-related diseases such as melanoma [14]. Notably, the issue of algorithmic bias extends beyond historically marginalized groups, potentially affecting anyone whose profile deviates from the predominant characteristics of the training datasets, whether in terms of skin color, gender, age, disease characteristics or even the hospital's zip code [12].

Nevertheless, instead of viewing these biases purely as flaws, they can be seen as "Data Artifacts" —records of societal values, healthcare practices, and historical inequities. By examining biased clinical data through this lens, researchers can uncover underlying patterns of exclusion and injustice that persist in healthcare. As proposed by Ferryman et al. (2023), this artifact-based approach can help AI developers not only detect and avoid bias, but also understand the root causes of health inequities. Such an understanding is crucial for medical research overall, and especially for developing AI systems that do not merely replicate existing injustices but actively contribute to more equitable healthcare practices [15]. By treating biased data as informative artifacts, we can examine healthcare data more holistically, uncovering population inequities and suggesting novel uses of AI to detect health equity–relevant data patterns.

Furthermore, there has been growing awareness on the need for transparency and accountability for AI medical applications. With it, the investigation of bias in AI algorithms and medical devices is a rapidly advancing field. Recently, the European Parliament, through the EU-AI Act [16], the White House, via the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [17], and the US Department of Health and Human Services, through the final rule on section 1557 of the Affordable Care Act, have initiated measures to mitigate bias in AI algorithms. However, much of the effort in this domain has been directed towards post hoc analysis—examining models for bias after their development and deployment [18]. We see this approach as costly, inefficient, and unable to promote systemic change.

Some researchers have put forward commendable efforts aimed at enhancing the understanding of datasets' collection processes, origins, development intents, recommended uses, and ethical considerations. These initiatives seek to establish standardized means for researchers and developers to quickly access critical information about datasets intended for training medical devices, algorithms, or conducting epidemiological research. Notable among these efforts are Data Cards [19], Data Statements [20], Datasheet for Datasets [21], Model Cards [22], AI-Usage Cards [23], and the Dataset Nutritional Label [18]. Each of these proposals contributes with valuable frameworks for documenting various aspects of datasets and models, facilitating a more responsible and informed use of data in AI development.

However, most existing initiatives primarily address general dataset characteristics and usage guidelines without delving deeper into the specific biases, or "artifacts" that datasets may contain. These existing frameworks, while foundational, are not equipped to dynamically track or update bias-related issues as they evolve or as new evidence comes to light. Moreover, they fall short of providing the nuanced understanding required to preemptively recognize and investigate biases specific to each dataset. This oversight underscores the critical need

for a community-based repository that systematically indexes, catalogs, and describes biases as informative data artifacts.

We propose the development of the *Data Artifacts Glossary*—a dynamic, open-source framework that serves as a collaborative platform for examining healthcare data bias as artifacts. By expanding the technical approach to data bias in AI development to include sociotechnical perspectives, the Glossary considers historical and current social contexts as important factors in addressing bias. This expanded approach not only aids in avoiding bias in model development but also serves the public health goal of understanding population inequities and suggests novel uses of AI to detect health equity–relevant data patterns.

Methods

The development of the *Data Artifacts Glossary* was guided by a methodology inspired by the Delphi method [24], a structured communication technique well-suited for achieving consensus among a diverse group of experts. Our team consisted of clinicians, computer scientists, data scientists, researchers, project managers, specialists in education, and legal experts. The process began with an initial round of discussions where each team member independently provided their insights on the potential sources of bias in healthcare datasets. These insights were compiled into a preliminary list of bias types and framework concepts. To ensure a comprehensive approach, we conducted an extensive literature review to identify existing frameworks and data documentation methods, including examining Data Cards, Data Statements, Datasheet for Datasets, Model Cards, AI-Usage Cards, and the Dataset Nutritional Label.

Following this, several rounds of structured discussions were conducted, each designed to refine and expand upon the preliminary concepts. These rounds involved feedback and structured iterations, facilitating a thorough examination and synthesis of diverse perspectives. Through this iterative process, the team reached a consensus on the most pertinent categories of bias to include in the glossary. Each category was chosen based on its relevance to clinical data and its potential impact on AI applications. Additionally, considerable discussion was dedicated to designing the structure of the glossary itself. Finally, the methodology included pilot testing the glossary with the MIMIC-IV dataset to validate its structure and utility. The pilot involved a detailed review of the dataset's published literature to identify and document specific biases, followed by the incorporation of this information into the glossary framework.

Results

The *Data Artifacts Glossary* is envisioned as a collaborative platform designed to systematically document and update biases associated with both public and non-public healthcare datasets. Unlike existing frameworks that provide static snapshots of data characteristics, this *Glossary* aims to establish a dynamic, community-driven repository where biases are continually identified, reported, and potential mitigation strategies revised. By viewing biased clinical data as informative artifacts, the *Data Artifacts Glossary* facilitates a deeper examination of societal values, healthcare practices, and historical inequities reflected in the data.

This living document will serve as a comprehensive reference point for researchers, clinicians, and AI developers, allowing them to understand not only the general attributes of a dataset but also the specific biases it may harbor. By integrating contributions from a diverse community of stakeholders, the *Data Artifacts Glossary* will evolve with the expanding landscape of medical data and emerging insights into biases, ensuring that the information remains current and relevant.

The *Data Artifacts Glossary* will adopt a collaborative model inspired by renowned open-source software practices, similar to those used by projects like Linux and Python. This approach will incorporate several key practices that have contributed to the success and widespread adoption of these software projects: Version Control, Public Reviews, and Documentation. First, the glossary will use a robust version control system at first facilitated by GitHub, a robust platform renowned for its strong collaborative features. This will allow multiple community members—including researchers, clinicians, AI developers, and other stakeholders—to simultaneously work on the glossary, efficiently tracking changes and managing versions. This transparent process ensures that every modification to the glossary's codebase is well-documented and accessible. Second, modifications and enhancements to the glossary will be handled through "pull requests". These requests, which community members can submit, are essentially proposals for revisions or additions to the glossary. Each pull request is made available publicly for review, fostering a rigorous peer review process. This process is managed by project maintainers who are selected based on their expertise and commitment to promoting unbiased AI in healthcare. The peer review ensures that all contributions adhere to high standards of quality and functionality.

Lastly, comprehensive documentation will be a cornerstone of the *Data Artifacts Glossary* project. Effective documentation is vital in open-source projects as it helps new users understand how to utilize the tool and aids new contributors in grasping the codebase and

the project's architecture. To allow contributions from researchers or community members who may not be familiar with coding, we introduced a more intuitive feature for collaboration. These users can access the *Data Artifacts Glossary* on GitHub via a provided link and submit suggestions through a user-friendly form (detailed in the "*Data Artifacts Glossary* Contribution Guide" tab). Developers can then review and incorporate these suggestions into the *Data Artifacts Glossary*, fostering broader participation and diverse input. Figure 1 provides a clear visual representation of the collaborative workflow and its components.

This methodology aims to foster an ongoing, dynamic update process and ensure that the glossary maintains a high level of academic rigor. The open-source model is designed to promote inclusivity and collective responsibility, essential for addressing the multifaceted nature of biases in healthcare datasets. By leveraging the collective intelligence of an interdisciplinary community, the *Glossary* facilitates the examination of biases as artifacts within their broader social and historical contexts, promoting a deeper understanding of the root causes of health inequities. This approach mirrors the principles of openness, peer review, and community engagement that are hallmarks of both academic rigor and the key practices of open-source projects.

The platform also features detailed documentation on each dataset, including its origin, collection process, and any amendments made to its associated biases, thereby providing transparency and traceability. In essence, the *Data Artifacts Glossary* will act as both a repository and a forum, fostering a collaborative environment for sharing knowledge and best practices in addressing dataset biases.

The ultimate goal of the *Data Artifacts Glossary* is to enhance the integrity and efficacy of AI applications in healthcare by providing a resource that helps mitigate the risk of bias from the beginning. By equipping stakeholders with detailed, up-to-date information on dataset biases, the Glossary aims to aid in the development of more accurate and fair AI algorithms. More importantly, by viewing biases as informative artifacts, the *Glossary*

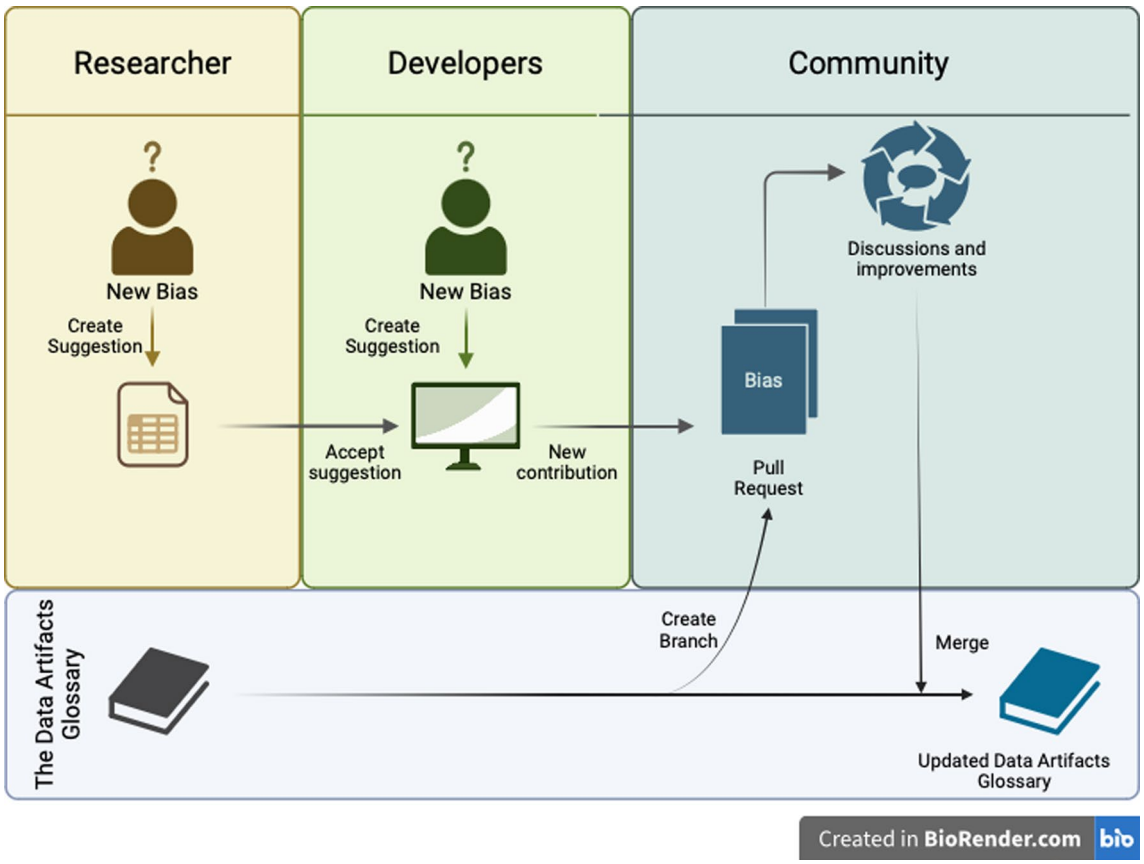


Fig. 1 Workflow diagram illustrating the collaborative process of the *Data Artifacts Glossary*. Researchers and developers can suggest new biases which are then reviewed and potentially accepted by the community. Once approved, these suggestions are merged back into the *Data Artifacts Glossary,* ensuring it remains an up-to-date and evolving resource

helps uncover and address the root causes of healthcare inequities, offering a more holistic approach to ethical AI use. It also aims to support the broader objective of ethical AI use, aligning with international efforts to ensure that AI systems are safe, secure, and trustworthy.

Suggested first version for MIMIC-IV

To demonstrate the practical application and utility of the *Data Artifacts Glossary*, we initiated a beta version of this platform (Figs. 2, 3, and 4) using the Medical Information Mart for Intensive Care (MIMIC-IV) dataset [25]. This dataset, consisting of de-identified health data associated with over seventy thousand patients admitted to critical care units of the Beth Israel Deaconess Medical Center in Boston, Massachusetts, is widely used for research in various domains of healthcare.

Here, we aim to suggest one potential structure as a starting point to populate the *Data Artifacts Glossary,* consisting of four initial categories, namely: Participants not missing at random, Validity of data points, Data not missing at random, and Miscellaneous. Of note, we do not aim to provide the final structure of the *Data Artifacts Glossary*, nor do we claim that the following categories are exhaustive.

Participants not missing at random

This category captures bias stemming from absence or underrepresentation of specific patient groups within the dataset, encompassing not only demographic factors but also clinical conditions, socioeconomic statuses, and accessibility variables which may skew research outcomes and subsequent clinical applications. The *Data Artifacts Glossary* under this category aims to illuminate the hidden disparities by documenting the absence of certain groups due to various selection biases or data collection constraints. This awareness is critical as it allows researchers and clinicians to critically evaluate the dataset and its applicability to the target patient population, ensuring that medical interventions developed from AI models do not inadvertently perpetuate health inequities. For example, a study using MIMIC-IV data found that Asian, Black, and Hispanic patients received invasive ventilation at significantly lower rates than White patients, despite presenting with similar clinical severity [26], indicating a potential systemic bias in treatment practices across racial lines. This discrepancy may be due to implicit biases in clinical decision-making, or differences in how symptoms are assessed and acted upon for patients of different ethnicities.

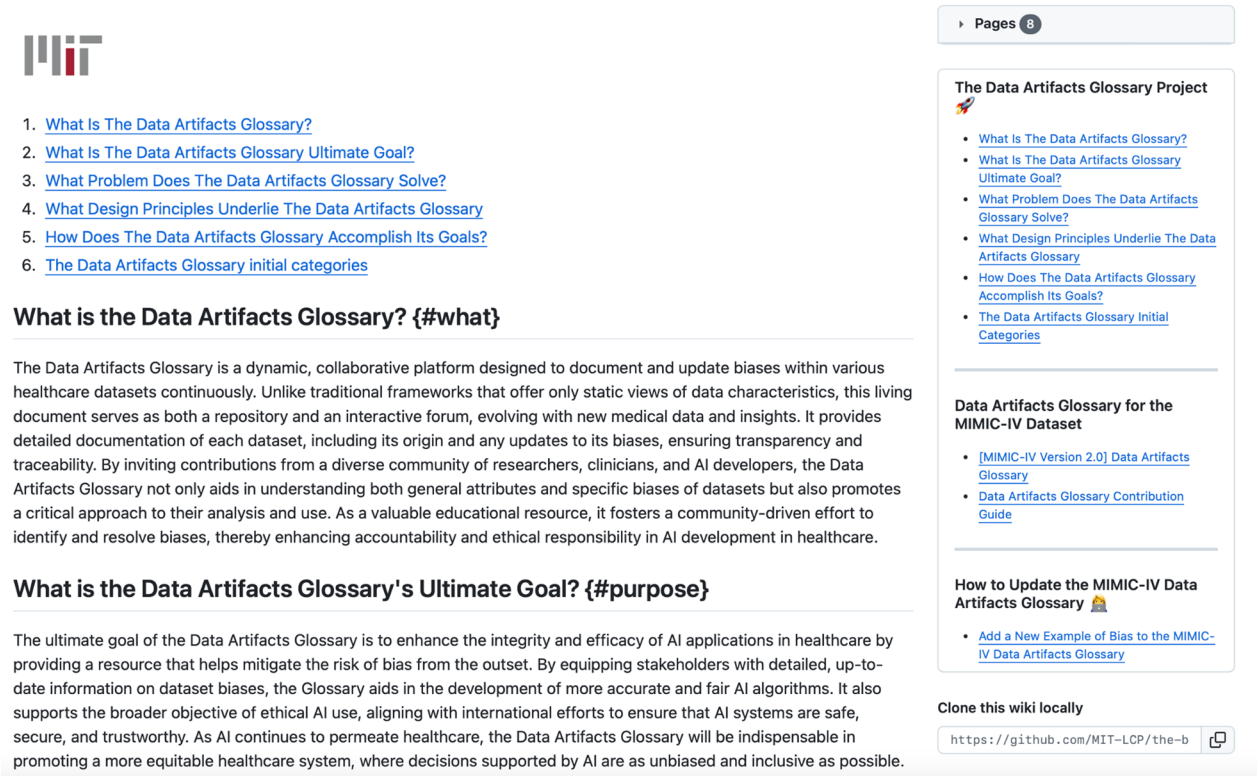
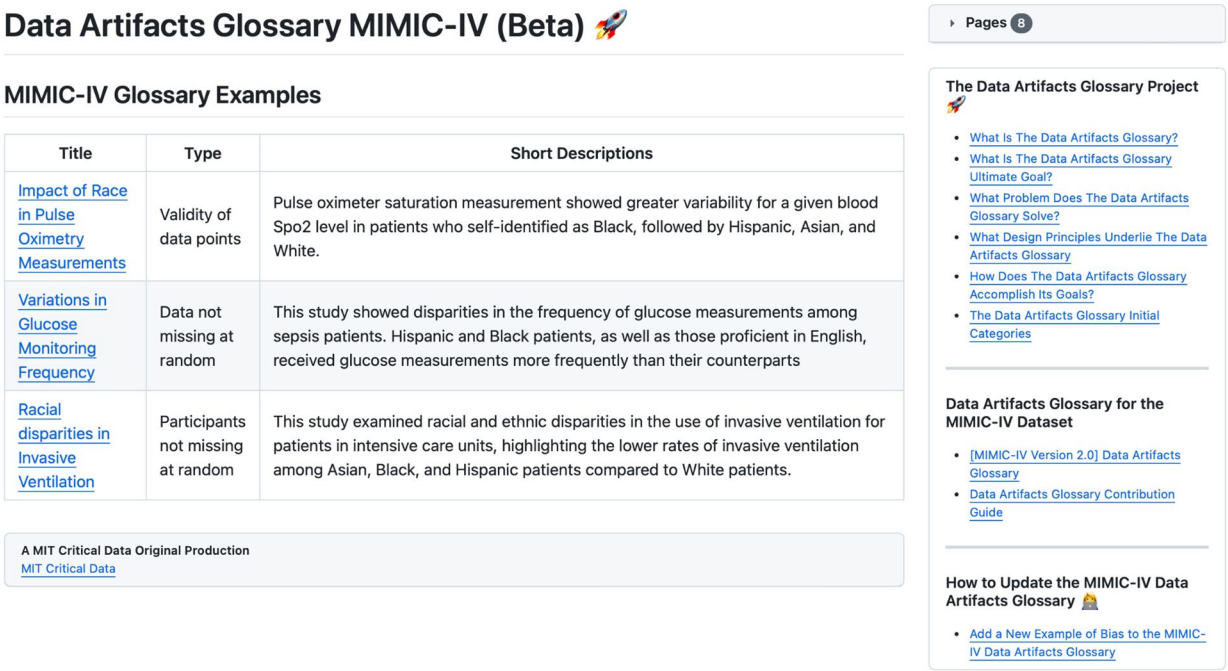


Fig. 2 Screenshot of the main page of the *Data Artifacts Glossary*. This page provides an overview of the project's goals, design principles, and initial categories, as well as links to detailed descriptions and guides for contributing to the *Data Artifacts Glossary* for the MIMIC-IV dataset



Data Artifacts Glossary MIMIC-IV (Beta) 🚀

MIMIC-IV Glossary Examples

Title	Type	Short Descriptions
Impact of Race in Pulse Oximetry Measurements	Validity of data points	Pulse oximeter saturation measurement showed greater variability for a given blood Spo2 level in patients who self-identified as Black, followed by Hispanic, Asian, and White.
Variations in Glucose Monitoring Frequency	Data not missing at random	This study showed disparities in the frequency of glucose measurements among sepsis patients. Hispanic and Black patients, as well as those proficient in English, received glucose measurements more frequently than their counterparts
Racial disparities in Invasive Ventilation	Participants not missing at random	This study examined racial and ethnic disparities in the use of invasive ventilation for patients in intensive care units, highlighting the lower rates of invasive ventilation among Asian, Black, and Hispanic patients compared to White patients.

A MIT Critical Data Original Production
MIT Critical Data

Pages 8

The Data Artifacts Glossary Project 🚀

- What Is The Data Artifacts Glossary?
- What Is The Data Artifacts Glossary Ultimate Goal?
- What Problem Does The Data Artifacts Glossary Solve?
- What Design Principles Underlie The Data Artifacts Glossary
- How Does The Data Artifacts Glossary Accomplish Its Goals?
- The Data Artifacts Glossary Initial Categories

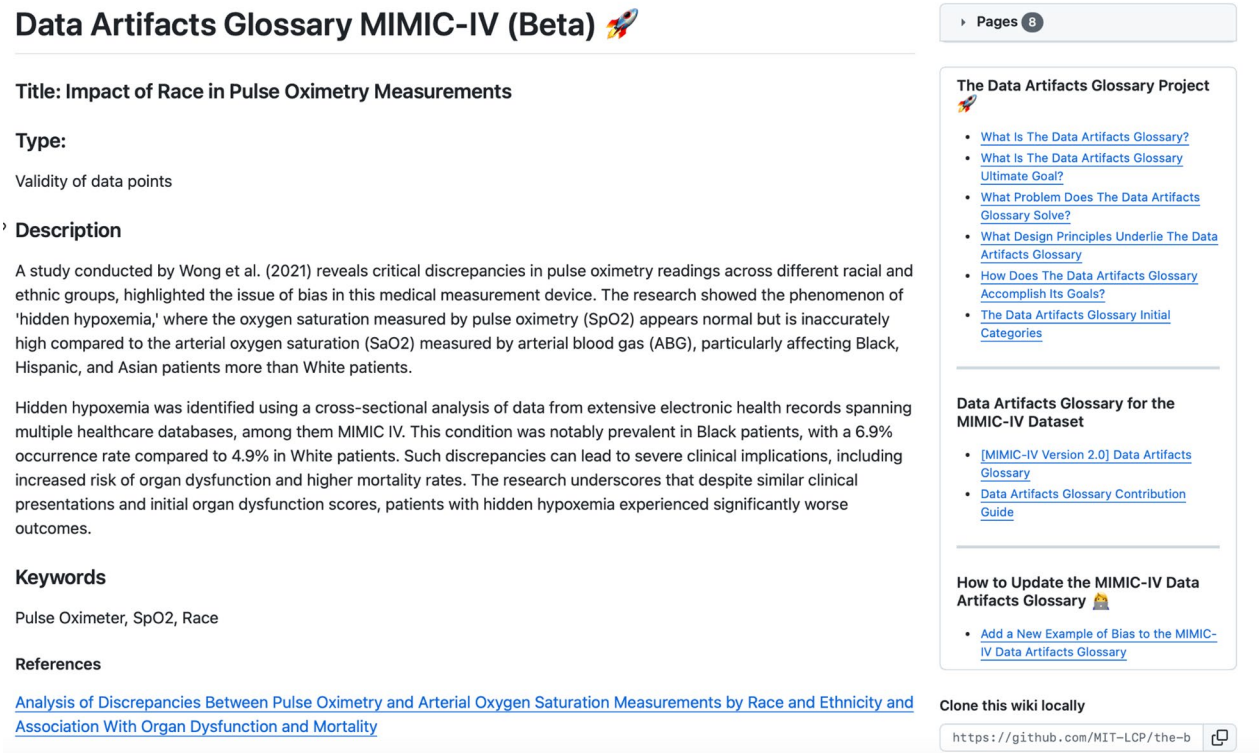
Data Artifacts Glossary for the MIMIC-IV Dataset

- [MIMIC-IV Version 2.0] Data Artifacts Glossary
- Data Artifacts Glossary Contribution Guide

How to Update the MIMIC-IV Data Artifacts Glossary 🧙

- Add a New Example of Bias to the MIMIC-IV Data Artifacts Glossary

Fig. 3 Screenshot of the Data Artifacts Glossary for the MIMIC-IV (Beta)



Data Artifacts Glossary MIMIC-IV (Beta) 🚀

Title: Impact of Race in Pulse Oximetry Measurements

Type:

Validity of data points

Description

A study conducted by Wong et al. (2021) reveals critical discrepancies in pulse oximetry readings across different racial and ethnic groups, highlighted the issue of bias in this medical measurement device. The research showed the phenomenon of 'hidden hypoxemia,' where the oxygen saturation measured by pulse oximetry (SpO2) appears normal but is inaccurately high compared to the arterial oxygen saturation (SaO2) measured by arterial blood gas (ABG), particularly affecting Black, Hispanic, and Asian patients more than White patients.

Hidden hypoxemia was identified using a cross-sectional analysis of data from extensive electronic health records spanning multiple healthcare databases, among them MIMIC IV. This condition was notably prevalent in Black patients, with a 6.9% occurrence rate compared to 4.9% in White patients. Such discrepancies can lead to severe clinical implications, including increased risk of organ dysfunction and higher mortality rates. The research underscores that despite similar clinical presentations and initial organ dysfunction scores, patients with hidden hypoxemia experienced significantly worse outcomes.

Keywords

Pulse Oximeter, SpO2, Race

References

Analysis of Discrepancies Between Pulse Oximetry and Arterial Oxygen Saturation Measurements by Race and Ethnicity and Association With Organ Dysfunction and Mortality

Pages 8

The Data Artifacts Glossary Project 🚀

- What Is The Data Artifacts Glossary?
- What Is The Data Artifacts Glossary Ultimate Goal?
- What Problem Does The Data Artifacts Glossary Solve?
- What Design Principles Underlie The Data Artifacts Glossary
- How Does The Data Artifacts Glossary Accomplish Its Goals?
- The Data Artifacts Glossary Initial Categories

Data Artifacts Glossary for the MIMIC-IV Dataset

- [MIMIC-IV Version 2.0] Data Artifacts Glossary
- Data Artifacts Glossary Contribution Guide

How to Update the MIMIC-IV Data Artifacts Glossary 🧙

- Add a New Example of Bias to the MIMIC-IV Data Artifacts Glossary

Clone this wiki locally

https://github.com/MIT-LCP/the-b

Fig. 4 Screenshot of detailed bias entry for the Data Artifacts Glossary for the MIMIC-IV (Beta)

Validity of data points

The second category examines the integrity of data collected, focusing on potential biases introduced through the use of various medical devices and data recording methodologies. This category is pivotal as it questions the foundational accuracy of the dataset itself —whether the data points reflect true patient states or are distorted by technological and procedural variances. By cataloging these potential sources of error, the *Data Artifacts Glossary* promotes a more nuanced understanding of the data, which is essential for developing reliable AI models. For instance, a study using MIMIC-IV found that hidden hypoxemia was more frequently under-detected in Black and Hispanic patients [27], underscoring the crucial bias in the accuracy of pulse oximetry measurements across different racial groups. This issue may have arisen because pulse oximeters were predominantly developed and calibrated using lighter-skinned populations, leading to decreased accuracy in individuals with darker skin pigmentation and underestimation of oxygen deprivation in these patients.

Data not missing at random

This category investigates the uneven data collection practices that may occur across various patient groups due to factors such as race, socioeconomic status, geographical location, and other demographic or contextual influences. It underscores the necessity to meticulously examine and question the consistency and fairness of data collection protocols and their execution among diverse patient populations. This detailed scrutiny is crucial for identifying and understanding the systemic errors and biases that could detrimentally impact clinical research and the training of AI algorithms. Under this category, the MIMIC-IV *Data Artifacts Glossary* lists the discrepancy between observed glucose measurement frequencies among different demographic groups, where significant increases in measurement frequency were found for individuals identified as male, Hispanic, Black, or English proficient [28]. One hypothesis is the presence of language barriers, making it more challenging for providers to communicate with non-English proficient patients to explain procedures, potentially leading to fewer glucose measurements being performed for these patients.

Miscellaneous biases

The fourth category encompasses a broad range of biases that do not neatly fit into the other categories but are nonetheless crucial for understanding and using the dataset responsibly. These might include biases related to the geographic location of data collection, time-period specific healthcare practices, or administrative biases in how data are recorded and processed. This section will be populated with examples that highlight impactful biases affecting data interpretation and application in AI systems.

The current glossary is not exhaustive and not static. New causes of bias having profound effects on downstream prediction, classification and optimization tasks will continuously be found for various datasets. From differential performance of medical devices used to measure physiologic signals across patient populations, to variation in the frequency of testing across patient populations that is not explained by clinical factors, to disparities in the performance of routine care that is typically assumed to be administered uniformly across patient populations. These discoveries are made possible by a collaborative community of users who are curating and analyzing its data and sharing those discoveries with each other. In the case of MIMIC, we expect the 70 k + users to contribute insights on sources of bias captured in the first *Data Artifacts Glossary*. We hope this serves as a lighthouse to establish more bias glossaries for other public and non-public datasets.

Discussion

The *Data Artifacts Glossary* represents a transformative approach to collaboratively identify and understand biases within healthcare datasets, not by merely viewing biases as flaws to be corrected, but by recognizing them as informative artifacts that reflect societal values, healthcare practices, and historical inequities. By fostering an environment where biases are continuously identified, documented, and addressed through community-driven efforts, this living document not only enhances the integrity of AI applications in healthcare but also promotes a more equitable healthcare system. The adoption of open-source principles and robust peer-review mechanisms ensures that the Glossary remains an up-to-date, transparent, and reliable resource, pivotal for developing AI tools that are both effective and fair. As we stand on the brink of a new era in healthcare, marked by technological advancements, the *Data Artifacts Glossary* serves as a crucial tool to ensure that these technologies benefit all segments of society equally, preventing the perpetuation of historical inequities.

The practical implementation of the *Data Artifacts Glossary* using the MIMIC-IV dataset demonstrates its significant potential in improving the quality of AI in healthcare. By meticulously categorizing and documenting specific biases within this widely-used dataset, the Glossary enables researchers and clinicians to better understand and address the inherent biases present in the data. This understanding is critical, as it allows for the identification and potential rectification of biases before

they influence AI models, which could otherwise perpetuate or even amplify existing health disparities. Using the MIMIC-IV *Data Artifacts Glossary* as an example, if O2 saturation is an important feature for developing a model, a researcher might choose to use arterial blood gas measurements instead of pulse oximetry. The *Data Artifacts Glossary* serves as both a reference and a guide, educating users about the various forms of bias and their implications, thus fostering a more informed and proactive approach to data management in healthcare AI.

Moreover, the collaborative nature of the *Data Artifacts Glossary* leverages the collective intelligence of a diverse group of stakeholders, including researchers, clinicians, and AI developers. This inclusive approach ensures that the Glossary remains comprehensive and reflects a wide range of perspectives, making it a robust tool for enhancing the fairness and accuracy of AI applications in healthcare. The community-driven contributions and rigorous peer review process ensure high-quality, reliable updates, keeping the Glossary relevant in a rapidly evolving field.

Limitations

Despite its strengths, the *Data Artifacts Glossary* is not without limitations. One of the primary challenges lies in ensuring widespread and consistent participation from the community. The quality and usefulness of the glossary depend heavily on the contributions of its users, which can be variable and influenced by individual biases and expertise levels. Additionally, while the glossary provides a framework for documenting biases and lists potential mitigation strategies, it cannot offer direct solutions to mitigate all sources of bias, requiring users to choose and apply their own methods and work-arounds. Addressing these limitations requires ongoing efforts to engage the community, streamline contributions, address sources of bias upon data generation, and perhaps develop additional tools and guidelines for bias mitigation.

Conclusion

The practical implementation of the *Data Artifacts Glossary*, demonstrated through the MIMIC-IV dataset, highlights its potential to significantly impact healthcare outcomes by providing a deeper, more nuanced understanding of dataset biases. This initiative is not merely a response to the growing complexity of medical datasets but a proactive measure to safeguard against the inadvertent introduction of biases by AI systems. By equipping researchers, clinicians, and policymakers with the knowledge to scrutinize and refine the datasets that train AI, the Glossary aids in the creation of more accurate and impartial medical AI applications. Furthermore, it serves as a novel tool for using AI to detect health equity–relevant data patterns, thereby expanding the potential of AI in promoting health equity. Moving forward, as the *Data Artifacts Glossary* continues to evolve, it will remain a vital resource for enhancing the fairness and accuracy of AI in healthcare, ensuring that it adapts to new challenges and insights in a rapidly advancing field.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12929-024-01106-6.

Additional file 1.

Acknowledgements
We would like to thank all contributors and reviewers who provided valuable insights and feedback, helping to refine and enhance the Bias Glossary framework.

Author contributions
All authors contributed to the writing of the manuscript and to the development of the concept for the Data Artifacts Glossary. RRG, NLW, and CMS drafted the initial version of the manuscript. SH, AEP, ART, IR and AIW conducted a review to identify examples of biases within the MIMIC-IV database. RRG, NLW and COF created the first version of the MIMIC-IV Data Artifacts Glossary. RRG, NLW and LAC conceived the original idea for the project..

Funding
The project did not receive funding.

Availability of data and materials
Github code and wiki: https://github.com/MIT-LCP/the-bias-glossary/wiki.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests. Dr. L. A. Celi is funded by the National Institute of Health through R01 EB017205, DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, and the National Science Foundation through ITEST #2148451. Dr. C. M. Sauer is supported by the German Research Foundation funded UMEA Clinician Scientist Program, under FU356/12-2.

Author details
[1]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [3]Laboratory for Clinical Research and Real-World Evidence, Department of Artificial Intelligence In Medicine, University Hospital Essen, Essen, Germany. [4]Department of Hematology and Stem Cell Transplantation, University Hospital Essen, Essen, Germany. [5]Division of Pulmonary, Allergy, and Critical Care Medicine, Duke University, Durham, NC, USA. [6]Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA, USA. [7]Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Department of Philosophy, Universitat Autónoma de Barcelona, Barcelona, Spain. [9]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA. [10]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA.

Received: 27 May 2024 Accepted: 17 November 2024
Published online: 04 February 2025

References

1. Basu K, Sinha R, Ong A, Basu T. Artificial intelligence: how is it changing medical sciences and its future? Indian J Dermatol. 2020;65(5):365–70. https://doi.org/10.4103/ijd.IJD_421_20.
2. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23(1):689. https://doi.org/10.1186/s12909-023-04698-z.
3. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. J Gen Intern Med. 2019;34(8):1626–30. https://doi.org/10.1007/s11606-019-05035-1.
4. Sauer CM, Chen LC, Hyland SL, Girbes A, Elbers P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. Lancet Digit Health. 2022;4(12):e893–8. https://doi.org/10.1016/S2589-7500(22)00154-6.
5. Pannucci CJ, Wilkins EG. Identifying and avoiding bias in research. Plast Reconstr Surg. 2010;126(2):619–25. https://doi.org/10.1097/PRS.0b013e3181de24bc.
6. Chapman EN, Kaatz A, Carnes M. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. J Gen Intern Med. 2013;28(11):1504–10. https://doi.org/10.1007/s11606-013-2441-1.
7. O'Sullivan ED, Schofield SJ. Cognitive bias in clinical medicine. J R Coll Physicians Edinb. 2018;48(3):225–32. https://doi.org/10.4997/JRCPE.2018.306.
8. Hoffman KM, Trawalter S, Axt JR, Oliver MN. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. Proc Natl Acad Sci. 2016;113(16):4296–301. https://doi.org/10.1073/pnas.1516047113.
9. Nsiah-Kumi P, Ortmeier SR, Brown AE. Disparities in diabetic retinopathy screening and disease for racial and ethnic minority populations–a literature review. J Natl Med Assoc. 2009;101(5):430–7. https://doi.org/10.1016/s0027-9684(15)30929-9.
10. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. J Am Med Inform Assoc JAMIA. 2020;27(12):2020–3. https://doi.org/10.1093/jamia/ocaa094.
11. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176–82. https://doi.org/10.1038/s41591-021-01595-0.
12. Saint James Aquino Y. Making decisions: Bias in artificial intelligence and data-driven diagnostic tools. Aust J Gen Pract. 2023;52(7):439–42. https://doi.org/10.31128/AJGP-12-22-6630.
13. Muzammil MA, Javid S, Afridi AK, et al. Artificial intelligence-enhanced electrocardiography for accurate diagnosis and management of cardiovascular diseases. J Electrocardiol. 2024;83:30–40. https://doi.org/10.1016/j.jelectrocard.2024.01.006.
14. Kamulegeya L, Bwanika J, Okello M, et al. Using artificial intelligence on dermatology conditions in Uganda: a case for diversity in training data sets for machine learning. Afr Health Sci. 2023;23(2):753–63. https://doi.org/10.4314/ahs.v23i2.86.
15. Ferryman K, Mackintosh M, Ghassemi M. Considering biased data as informative artifacts in AI-assisted health care. N Engl J Med. 2023;389(9):833–8. https://doi.org/10.1056/NEJMra2214964.
16. EU AI Act: first regulation on artificial intelligence | Topics | European Parliament. 11:40:00.0. Accessed February 12, 2024. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.
17. House TW. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The White House. October 30, 2023. Accessed February 12, 2024. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
18. Holland S, Hosny A, Newman S, Joseph J, Chmielinski K. The dataset nutrition label: a framework to drive higher data quality standards. Published online May 9, 2018. https://doi.org/10.48550/arXiv.1805.03677.
19. Pushkarna M, Zaldivar A, Kjartansson O. Data cards: purposeful and transparent dataset documentation for responsible AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT'22. Association for Computing Machinery; 2022:1776–1826. https://doi.org/10.1145/3531146.3533231.
20. Bender EM, Friedman B. Data statements for natural language processing: toward mitigating system bias and enabling better science. Trans Assoc Comput Linguist. 2018;6:587–604. https://doi.org/10.1162/tacl_a_00041.
21. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for Datasets. Published online December 1, 2021. https://doi.org/10.48550/arXiv.1803.09010.
22. Donald A, Galanopoulos A, Curry E, et al. Towards a semantic approach for linked dataspace, model and data cards. In: Companion Proceedings of the ACM Web Conference 2023. WWW '23 Companion. Association for Computing Machinery; 2023:1468–1473. https://doi.org/10.1145/3543873.3587659.
23. Wahle JP, Ruas T, Mohammad SM, Meuschke N, Gipp B. AI usage cards: responsibly reporting AI-generated Content. Published online May 9, 2023. https://doi.org/10.48550/arXiv.2303.03886
24. Shang Z. Use of Delphi in health sciences research: a narrative review. Medicine (Baltimore). 2023;102(7): e32829. https://doi.org/10.1097/MD.0000000000032829.
25. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1. https://doi.org/10.1038/s41597-022-01899-x.
26. Abdelmalek FM, Angriman F, Moore J, et al. Association between patient race and ethnicity and use of invasive ventilation in the United States. Ann Am Thorac Soc. 2024;21(2):287–95. https://doi.org/10.1513/AnnalsATS.202305-485OC.
27. Wong AKI, Charpignon M, Kim H, et al. Analysis of discrepancies between pulse oximetry and arterial oxygen saturation measurements by race and ethnicity and association with organ dysfunction and mortality. JAMA Netw Open. 2021;4(11): e2131674. https://doi.org/10.1001/jamanetworkopen.2021.31674.
28. Teotia K, Jia Y, Link Woite N, Celi LA, Matos J, Struja T. Variation in monitoring: glucose measurement in the ICU as a case study to preempt spurious correlations. J Biomed Inform. 2024;153: 104643. https://doi.org/10.1016/j.jbi.2024.104643.

Publisher's Note