TITLE PAGE

Article Type:

Review

Title:

Moving Beyond the Empty Cell: The Threat of Decontextualized Healthcare Data

Author List:

Aya El Mir¹, Eric Bezerra de Sousa², Ignacio Mesina-Estarrón3, Leo Anthony Celi⁴,⁵,⁶, Moad Hani7, Mohammed Benjelloun7, Neha Nageswaran8, Saïd Mahmoudi7, Shaheen Siddiqui9, Sreeram Sadasivam¹⁰, William Greig Mitchell11

Affiliations:

¹ New York University Abu Dhabi, Department of Engineering, Abu Dhabi, United Arab Emirates  
 ² Faculty of Medicine, University of Sao Paulo, Sao Paulo, Brazil 3 Department of Neurosurgery, Children’s Hospital of Orange County, Orange, CA, USA  
 ⁴ Laboratory for Computational Physiology, Massachusetts Institute of Technology, MA, USA  
 ⁵ Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA  
 ⁶ Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA  
 ⁷ Department of Computer Engineering and Management, Faculty of Engineering, University of Mons (UMONS), 7000 Mons, Belgium  
 ⁸ Robotic Assistant Systems, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany   
 ⁹ GeneSilico, Inc., Texas, USA   
 ¹⁰ EGYM SE, Munich, Germany  
 ¹¹ Department of Ophthalmology, Cambridge University Hospitals, Cambridge, UK

*Authors are listed alphabetically. All authors contributed equally to the best of their abilities.*

Word count:

3,762 words

Competing interests:

No competing interests to declare

Acknowledgements:

AI-assisted tools were used to improve the clarity and coherence of certain sections of this manuscript. Their contribution was limited to refinement of language wherein it was ensured that no distortion of the research content or data interpretation was made.

ABSTRACT

Missing, inaccurate or poorly documented data in healthcare is often treated as a technical problem to be statistically resolved via imputation, deletion, or modeling assumptions about randomness. However, such inaccuracies relate to far more complex socioeconomic and geopolitical issues, rather than ‘errors of data entry’ to be ameliorated with statistical modelling techniques. We outline that what is really missing or inaccurate is the *context* in which the data is collected – and that only by understanding this context can we begin to prevent artificial-intelligence’s (AIs) amplification of misleading, decontextualizeddata. We critically examine how traditional modelling methods fail to account for the factors that influence what data gets recorded, and for whom. We show how AI systems trained on decontextualizeddata reinforce health inequities at scale. And we review recent literature on context-aware approaches to understanding data, that incorporate metadata, social determinants of health, fairness constraints, and participatory governance to build more ethical and representative systems. Our analysis urges the AI and healthcare communities to move beyond the traditional emphasis on statistical convenience, toward socially grounded and interdisciplinary strategies for handling decontextualizeddata.

AUTHOR SUMMARY:

Healthcare data that is missing, incomplete, or inaccurately documented is often treated as a technical problem to be solved with statistical methods. We emphasize that this perspective overlooks the real issue: the data has been stripped of its context. Missing, incomplete or inaccurate data (collectively termed *decontextualized data*) is not random; it is shaped by human decisions, social barriers, and systemic inequalities.

Decontextualized healthcare data becomes increasingly dangerous as the use of AI in healthcare proliferates. Models trained on decontextualized data learn existing distortions as if objective truths. Consequently, their predictions risk reinforcing the very inequities that caused the flawed data in the first place and exacerbating health disparities at scale.

We argue for a paradigm shift towards understanding why data becomes decontextualized. This requires a concerted effort between machine learning communities and domain experts who understand data context. It is only through this partnership that we can begin to build models that account for the complex realities embedded in decontextualized healthcare data that cannot be solved by sophisticated modelling techniques alone.

MANUSCRIPT

### Introduction

Statistical techniques to manage missing healthcare data like imputation and deletion have been extensively researched and applied for decades, on the premise that data is missing either completely at random (MCAR), at random (MAR), or not at random (MNAR) (1, 2). These techniques, however, ignore an incontrovertible truth about missing healthcare data: that it is essentially *never* missing at random. What is really missing is an understanding of the *context* of the data. Rather than representing an empty cell to be imputed, deleted, or filled for better statistical model performance, missing healthcare data is the result of long-standing, intertwining geopolitical and socio-economic complexities, which demand far more nuanced and insightful approaches than those currently used if we are to meaningfully interpret them.

When approaching missing data in healthcare, statistical methods rarely appreciate how the data came about – how data was collected (or not collected) and by whom, the socio-economic policies or habits facilitating data collection (or not), the devices used to capture certain data signals (or mis-capture/not capture signals), and how final datasets are curated and made available for analysis (3).

Data may be missing because of socioeconomic barriers to documentation. For instance, in intensive care units (ICU), overnight blood sugar level (BSL) data may not be documented due to language discordance between staff and patient, necessitating an overnight interpreter and extra work – it’s easier to let the non-English speaking patient sleep without checking their BSL, rather than wake them and concurrently find a translator (4). Similarly, overnight bed-ulcer status might not be documented because a patient is too obese to roll with skeleton night staff, so it’s decided to wait till the morning. These data aren’t just ‘missing’; they’re the result of nocive social barriers prohibiting data collection.

Recorded data may be so inaccurate that real underlying information is cloaked by the erroneous reading. For instance, pulse oximetry data is much more inaccurate for patients with heavily pigmented skin (5), with falsely-high readings leading to systemic under-treatment and worse outcomes (6). Again, missing data context and subsequent inaccuracy here is not random, but indicative of inequalities in who equipment calibration and data collection has been optimized for (non-pigmented patients) versus who data is being collected from, in a clinical setting.

Data may not even make it to the point of collection in the first place – a form of ‘invisible’ missing data. Black patients suffering out of hospital cardiac arrest (OOHCA) are less likely to be resuscitated and subsequently transported to a medical facility (7), making their data less likely to be captured. Aboriginal & Torres Strait Islander Australians (hereafter Indigenous Australians) are more likely to self-discharge from ICU against medical advice, similarly making their critical illness data impossible to capture in the first place (8). And female patients are known to be more likely to have acute coronary syndrome (ACS) misdiagnosed by male physicians (9), making subsequent cardiovascular data less likely to be recorded.

When AI models are built with decontextualized data, they absorb these blind spots to generate flawed predictions at scale: the well described “garbage in, garbage out” AI phenomenon (10). While statistical methods that more accurately address contextual complexities embedded in data may exist, they cannot be developed by those who do not understand the data context at all points in the pipeline, from collection, documentation, curation, storage and distribution (3). Current statistical modelling methods to handle missing, inaccurate, and poorly recorded data (hereafter referred to collectively as ‘decontextualized data’) dismiss the context dictating why it is flawed and seek only to improve model accuracy parameters. Machine Learning communities working in isolation lack an understanding of the geopolitical and socio-economic context of such missing data, and are unable to develop such statistical tools alone. Without this understanding, gaping contextual cavities in data will continue to exist, posing most risk to those who’s data is most contextually-flawed – historically those already disadvantaged by current healthcare structures (11).

This paper is a call to collaborative action between those who understand data context and those developing statistical techniques to resolve decontextualizeddata. We propose solutions and methodological improvements for tackling decontextualizeddata so that we might stand a chance at developing insightful methods to minimize its impact on already marginalized populations.

### Rethinking Missing Data: The Importance of Context and Critique of Traditional Approaches

Understanding mechanisms that perpetuate decontextualizeddatarequires examining the structural factors underlying data collection. The data lifecycle can be divided into different pipeline stages, to the point of acquisition and use by statisticians, to help highlight geopolitical and socio-economical contextual issues.

1. Data Availability: Historically Disadvantaged Populations are Less Likely to Have Their Data Collected – ‘Invisible’ Missingness Contextualization.

Healthcare data may be unavailable to document in the first place, a form of ‘invisible’ missing data. This data is traditionally representative of patients from lower socioeconomic backgrounds, racial or ethnic minority groups, or communities with fragmented healthcare services across multiple regions. These are the populations whose healthcare stands most to gain by the sensitive use of AI, but will conversely become relatively more disadvantaged without adequate contextual understanding of their healthcare data.

The health-related disadvantage of Indigenous Australians is extensively described, with substantially higher rates of chronic disease and difficulty accessing primary or tertiary healthcare (12-16). Self-discharge rates against medical advice for Indigenous Australians are the highest in the world (17); indeed, they’re almost four-times more likely to self-discharge during an ICU admission than non-Indigenous Australians (8). Whilst the reasons for self-discharge are complex, they likely reflect an unfavorable perception of the unfamiliar hospital environment, perceived differences in care, institutionalized racism and/or communication barriers (8, 18, 19). Those who self-discharge against medical advice have almost 50% higher 8-year mortality compared to those who do not self-discharge against medical advice (adjusted hazard ratio 1.46; 95% confidence interval 1.01–2.1) (8). Despite being a particularly high-risk group, healthcare data for patients who self-discharge against medical advice becomes impossible to document, and its omission from morbidity/mortality prediction models means actual predictions underestimate how unwell these patients are.

‘Invisible’ missing data similarly applies to Black patients suffering OOHCA in the United States (US). Black patients experience significantly higher rates of OOHCA than White patients (20) and in rural US settings are up to five times less likely to survive OOHCA compared to patients living in more affluent areas of the country (21). Survival discrepancies following OOHCA are *not* due to biological variation between Black vs White patients – but contextual factors: lower likelihood of OOHCA being witnessed by bystanders, lower likelihood of bystander cardiopulmonary resuscitation if witnessed, lower availability of automatic electronic defibrillators, longer Emergency Response Service times and lower likelihood of early advanced airway management upon their arrival (7); leading to lower rates of hospitalization after OOHCA. Further, those that do make it to hospital are less likely to receive urgent interventions correlated with survival, like targeted temperature management, coronary angiogram or cardiac catheterization due to variability in hospital capabilities (i.e., 24/7 cardiac catheter labs) and standardized post-arrest pathways (7, 21) – further contextualizing why survival rates are poorer. Decontextualized data used to train models to predict survival after OOHCA for Black patients omit the data for those who did not make it in the first place, and for those who do, may erroneously predict substantially higher risk of mortality attributed toward their race, rather than the context of where data was collected i.e., in lower capability healthcare settings. Without un understanding this context, lifesaving interventions may be withdrawn earlier on the assumption that their outcome will be poor due to their race.

When certain populations are systematically underrepresented in datasets, such as Indigenous Australians or Black patients in the US, predictive models can also underestimate future healthcare needs. By failing to understand the data context, models misinterpret that relatively lower historical healthcare expenditure (which is actually due to relatively under-reported healthcare data) equates to lower future healthcare needs than are actually required – as has been demonstrated by Obermeyer et al (22).

1. Data Collection and Documentation: Incomplete, Inaccurate or Misleading Documentation due to Clinician Discretion or Workforce Variability – Clinical Environment Contextualization.

Even when recorded, data is often incomplete, inconsistent or misleading due to contextual conditions under which measurements are recorded. For example, the frequency of BSL measurements in ICU varies between patients depending on physician judgement, anxiety, and perceived risk of hypoglycaemia. When patients are perceived by clinicians as more at risk of hypoglycaemia, possibly due to clinician anxiety or perceived (rather than actual) risk, they are more likely to have more frequent BSL measurements. A recent review of the MIMIC-IV dataset analyzing almost 25,000 patients suggested as much, demonstrating that Black and Hispanic patients underwent BSL measurements 6% and 11% more frequently than White patients respectively, even when fully adjusted for illness severity and comorbidities (95% incidence rate ratio confidence intervals 1.01-1.12, and 1.01-1.21, respectively) (23). When context is ignored, spurious conclusions may be drawn i.e., that frequent BSL monitoring is linked to higher illness severity and poorer outcomes, when in reality the reverse causal direction is true, and may actually be more reflective of perceived risk of illness severity and BSL fluctuation, or clinician anxiety (23).

Staffing and shift models in ICU also affect documentation. Workflow interruptions (i.e., staff handover and shift changes) (23) and language barriers (i.e., language discordance between nursing staff and patients) make BSL recording less likely during those periods (4). The same publication using MIMIC IV data demonstrated that English-speaking patients are 8% more likely to have BSL measurements taken in US ICU’s vs non-English speaking patients (95% incidence rate ratio confidence intervals 1.01-1.15), possibly because they’re more likely to “speak up” without language discordance barriers (23). When such decontextualized data is analyzed, conclusions that certain patients had more frequent readings due to higher illness severity (rather than language concordance) are likely to be drawn. This is additionally concerning, given that more frequent BSL monitoring is associated with lower BSL fluctuations, hospital morbidity/mortality, and shorter length of stay (24-27), compounded further by the well-documented higher ICU mortality for non-English speaking patients without interpreters present, independent of frequency of BSL measurements (28, 29).

1. Data Saving and Storage: Understanding the Context Behind the Capacity for Consistent, Comprehensive Data Storage.

Understanding the context of variation in data storage capacities between healthcare facilities is equally critical. The availability of extensive datasets and significant computational power is an essential precursor to building robust, equitable models (11). Regional and institutional disparity means those with advanced technological infrastructure and large-scale data storage capabilities, such as affluent regions of the US (home to over 40% of the world’s databases alone), are overrepresented in the data used to train predictive models (11). Consequently, while models may perform well in environments where data originated, like the US, they lack generalizability and could perpetuate healthcare inequities when applied to populations from data-poor regions, like the Global South (30). Even within affluent, data rich areas with large-scale data storage capacity like the US, socially, ethnically or racially marginalized groups are more likely to have fragmented care across multiple institutions, and lower internet access/literacy preventing the use of online portals and patient-reported outcomes (31). This worsens scattered, incomplete data documentation for already-marginalized groups within affluent regions, similarly making model findings less generalizable to them. Ignoring the context of data provenance, and limitations imposed by heterogeneous data storage capacities poses further risk to model bias.

1. How Current Statistical Approaches Fail to Recognize Healthcare Data Context.

Most machine learning pipelines continue to rely on traditional statistical approaches focused solely on data completion. Techniques such as mean, median, and mode imputation; k-nearest neighbors (KNN); regression-based methods; forward and backward filling; interpolation; multiple imputation (including MICE); and model-based techniques like Kalman filters and mixed-effects models remain widespread (32-34). These methods, while mathematically sound, fall short in addressing contextual truths embedded in data. They typically assume that missingness can be explained within the data itself, often under the MAR or MCAR paradigms, assumptions that are not only overly optimistic but frequently invalid, ignoring the interplay of socioeconomic and geopolitical factors influencing missing data as described above.

For example, KNN and regression imputation assume similarity or linearity, which fails in marginalized or underrepresented populations where data is systematically lacking. Techniques like multiple imputation may provide robustness but are computationally intensive and still rely on the flawed assumption that missingness is ignorable if modeled well enough. These approaches prioritize statistical completeness over epistemological clarity. They risk creating an illusion of data integrity while masking the very structural barriers that created the missingness in the first place. Without this awareness of the importance of data context, algorithmic outputs risk reproducing the very inequities they seek to mitigate.

1. Epistemology: Acknowledging Our Own Limitations.

Recognizing missing context requires epistemic humility: acknowledgment that our understanding of data is shaped and limited by our own individual experiences and exposures. Data analysis often proceeds as though those interpreting the data possess full knowledge of its origins. Before any exploratory analysis begins, those responsible for data analysis and modelling i.e., the machine learning community must ask “how did this data come about?”; acknowledging that the social and structural conditions surrounding data generation are as important as the data itself.

### Context-Aware Solutions: Moving Beyond Traditional Paradigms, Toward Context-Aware Strategies to Promote Equity, Transparency and Accountability in AI.

While prior sections have laid a comprehensive foundation on the socio-political dimensions of missing data, it is imperative that we extend the conversation toward actionable, technologically informed, and ethically conscious approaches rooted in context-aware design.

1. Context-Enriched Metadata Collection and System Design

Research in context-aware computing and ambient intelligence has shown that capturing metadata dynamically, in real-time and across varying environmental and social conditions, yields more accurate and equitable data representations (35, 36). To move beyond decontextualized variables, data systems must routinely record situational factors such as time, location, language spoken, personnel present, and institutional pressure points like understaffing during night shifts (37, 38). Equally important is the documentation of those collecting the data. The training, cultural background, and experience of healthcare workers can introduce latent biases into the dataset, biases that remain invisible unless explicitly recorded and analyzed (37, 39). Addressing these sources of error begins with workforce preparedness: training programs that incorporate cultural competency, linguistic sensitivity, and bias mitigation have been shown to improve data fidelity across diverse populations (11). Metadata frameworks must also capture environmental and temporal variables, including day of the week, shift time, and geographic setting, which can significantly affect the quality and interpretation of physiological data. For instance, studies have shown that readings collected during emergency situations can vary widely depending on time and contextual stressors (22, 40) i.e., more frequent BSL measurements during a period of clinical instability for an ICU patient, as discussed in section 2 (23). Without this context, such variability risks being misinterpreted as clinical noise rather than a reflection of underlying system dynamics.

1. Embedding Social Determinants in Imputation and Machine Learning Pipelines

To address the limitations of traditional approaches, context-aware imputation strategies must move beyond purely statistical assumptions about randomness. Standard methods often ignore the social structures that drive missingness or inaccuracy. For example, if blood pressure measurements are systematically absent for rural women due to gendered norms around medical visits, filling those gaps with population averages risks introducing epistemic violence, further marginalizing the very groups already underserved (41, 42). Instead, imputation models must explicitly incorporate the social and infrastructural factors shaping data collection. Multi-level models that account for gender, region, language, and infrastructure variability offer a more grounded approach to estimating missing values (43, 44). These models can be further strengthened through ensemble techniques that combine stratified imputation, hierarchical modeling, and institutional metadata weighting, together enabling more robust and context-sensitive estimations (45, 46). In parallel, generative modeling methods can benefit from contextual inputs as well. Including attributes such as primary language, insurance status, and socio-economic indicators has shown promise in improving the quality of synthetic data while minimizing the risk of reproducing and amplifying existing biases (47, 48).

1. Fairness Through Architectural Alignment

Fairness in machine learning cannot be retrofitted; it must be embedded into system design from the outset. While techniques like adversarial debiasing, a method designed to mitigate unwanted biases in machine learning models by leveraging adversarial learning techniques, offer promise – their effectiveness depends on being guided by domain-specific fairness constraints i.e., fairness definitions that make sense for clinical decision-making. In healthcare, this requires nuanced trade-offs, such as balancing demographic parity with clinical urgency to ensure that fairness metrics do not come at the expense of patient safety – to avoid an algorithm treating everyone identically, at the risk of one who needs more urgent care (22, 45). Crucially, fairness is not universal; it is context-dependent. What constitutes a fair outcome in urban Canada may diverge significantly from what is considered fair in rural Morocco. Cultural norms, resource availability, and local health priorities must inform how fairness is defined and operationalized (49). Designing for fairness also requires transparency. Model explainability layers that incorporate contextual attributes can help users and stakeholders understand how predictions are made and where potential biases may arise. This is especially critical in historically under-represented communities like the Global South, Black, and Indigenous populations, which often experience data extraction without benefit – like the well described “parachute research” phenomenon where data is taken from these communities for the benefit of more economically-developed, data-rich countries and populations – creating an erosion of trust in technological systems (3, 39, 50).

1. Federated and Participatory Frameworks for Ethical Generalization

Federated learning offers a solution to the persistent tension between data privacy and model generalization. In federated learning, models are trained across decentralized data sources, such as hospitals or clinics, without transferring the underlying data. Instead, only model updates are shared and aggregated, allowing institutions to retain control over their data while still contributing to a collective, robust algorithm (51). When carefully tuned to respect institutional context, federated models can improve generalizability across diverse settings without compromising patient confidentiality. To make this collaboration possible, shared ontologies and documentation standards are essential. Ontologies provide a structured way to define and relate concepts, such as symptoms, diagnoses, or treatment protocols, ensuring that data from different sources can be interpreted consistently. Without such alignment, semantic mismatches across institutions can undermine the integrity of cross-institutional learning (52). Moreover, the inclusion of patients and community actors in data governance, via participatory design or data stewardship councils, can redirect the power dynamics behind health data use (53). Tools like data trust scorecards and consent granularity interfaces should be implemented to uphold data agency, especially for historically underserved populations. Indigenous Data Sovereignty movements offer transformative models through CARE Principles (Collective benefit, Authority to control, Responsibility, Ethics) (54). Evidence shows Indigenous-controlled initiatives achieve more accurate and complete data (55), better health outcomes through culturally appropriate data use (56), increased community participation (57), and high response rates (~90%) under OCAP principles (58).

1. Rethinking Risk, Accountability, and Regulation

The path to ethical and effective context-aware systems is paved with regulatory foresight. Initiatives such as the CONSORT-AI guidelines underscore the need for clinical and technical stakeholders to co-develop auditing criteria for AI models (59). Institutions should implement dynamic audit trails, bias impact logs, and socio-technical risk registries that evolve with the data landscape. Differential privacy techniques can further support these efforts, enabling secure sharing of metadata about missingness without compromising patient confidentiality (38, 60). Risk assessment frameworks should incorporate not only statistical risks but also social harms related to representational imbalance and demographic exclusion (61, 62). To achieve sustainable reform, missing data handling must be repositioned not as a peripheral statistical problem, but as a core ethical design concern – one that reflects our commitment to justice, representation, and responsible technology (61, 63).

1. Institutional Incentives, Open-Source Ecosystems, and Pilot Integration

A frequently overlooked dimension of implementation is the incentive structure required to make context-aware data practices viable. Institutions should align incentives with outcomes that reward improved equity, documentation quality, and bias reduction (64, 65). These could include accreditation credits, targeted funding opportunities, or increased research visibility tied to transparent data stewardship and inclusive design. Scaling context-aware practices also depends on accessible and adaptable technical infrastructure. Open-source machine learning libraries, such as TensorFlow and PyTorch, which are widely used for building deep learning models, and scikit-learn, a foundational library for classical machine learning tasks, offer flexible frameworks for experimentation and deployment. These tools, when extended with plug-ins or modules that support context tracking and metadata capture, can lower the barriers to implementing fairness-aware and provenance-conscious models across healthcare systems (66). Similarly, cloud-based healthcare APIs can be pre-configured to support compliance requirements and automatically log contextual variables such as collection setting, practitioner role, or equipment used (51).

Pilot programs, small-scale, controlled deployments of new tools or approaches, remain a critical lever for institutional change. Rather than pursuing broad, unfocused implementation, targeted pilots should concentrate on data pipelines that are known to concentrate bias and inequality, such as emergency room triage, algorithmic risk scoring, or diagnostic tools reliant on language fluency. These high-stakes, bias-prone settings provide an opportunity to test the real-world benefits of context-aware strategies, evaluate their effectiveness, and build institutional trust. Demonstrating tangible improvements in these contexts can catalyze broader policy support and system-wide adoption (67). Comprehensive quality improvement programs achieved remarkable results, with one hospital improving data quality scores from 41% to 77% in three months, reaching 80% thereafter (68). Successful interventions included data quality reporting and personalized feedback (61%), IT-related solutions (54%), and comprehensive training (44%) (69).

1. Cross-Industry Implementation Models

Healthcare's endemic data quality failures, ranking among the worst-performing sectors for completeness, demand systematic adoption of proven cross-industry methodologies (70).

Manufacturing's Quality 4.0 offers real-time frameworks through automated sensor networks with real-time data ‘detection and correction’ capabilities (71), predictive maintenance systems that proactively identify data quality issues, and machine learning integration with traditional quality management which could be used for continuously monitoring patient data in clinical healthcare environments (72, 73). Financial sector approaches provide Principal Component Analysis (PCA) for dimensionality reduction in missing data scenarios, cross-sectional and time-series hybrid models outperforming traditional imputation methods, and automated fraud detection systems handling missing data patterns in real-time, which may equally have applicability in clinical data management. (74, 75). Aviation's safety-first management achieved 60% reduction in safety-related incidents through integrated reservation, flight management, customer relationship, and operational systems with predictive maintenance reducing downtime by 40% (73, 76). Technology's IoT sector revolutionized quality through automated validation systems identifying low-quality data immediately, edge computing processing data locally before transmission, and machine learning for predictive data quality assessment (58, 71, 77).

### Conclusion

No amount of algorithmic refinement or sophisticated modelling techniques can compensate for datasets that fundamentally fail to capture the context of their own creation. Throughout this paper, we have examined the structural, epistemic, and algorithmic dimensions of decontextualized data, emphasizing that as AI becomes more deeply embedded in clinical decision-making, the risks of training models on such data grow increasingly urgent.

Rather than aiming for mathematical accuracy, we must foster a paradigm shift that moves toward deep, interdisciplinary collaboration. We must recognize missing, incomplete and inaccurate data not as a statistical nuisance, but an ethical and systemic signal. This requires a concerted effort between the machine learning community and domain experts; clinicians, healthcare workers, and community representatives, who understand data context. It is only through this partnership that we can begin to build models that account for the complex realities and what data is trying to communicate.

Ultimately, we must reimagine AI systems not simply as tools of optimization, but as opportunities for repair, systems that do not obscure what is decontextualized, but illuminate it, giving voice to those historically left out, and shaping a more equitable and accountable future.

REFERENCES:

1. Pham TM, Pandis N, White IR. Missing data: Issues, concepts, methods. Seminars in Orthodontics. 2024;30(1):37–44.

2. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. Journal of Big Data. 2021;8(1).

3. Mitchell WG, Wawira JG, Celi LA. Rebooting artificial intelligence for health. PLOS Global Public Health. 2025;5(1):e0004171.

4. Twersky SE, Jefferson R, Garcia-Ortiz L, Williams E, Pina C. The Impact of Limited English Proficiency on Healthcare Access and Outcomes in the U.S.: A Scoping Review. Healthcare. 2024;12(3):364.

5. Kyriacou PA, Charlton PH, Al-Halawani R, Shelley KH. Inaccuracy of pulse oximetry with dark skin pigmentation: clinical implications and need for improvement. British Journal of Anaesthesia. 2023;130(1):e33–e6.

6. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial Bias in Pulse Oximetry Measurement. New England Journal of Medicine. 2020;383(25):2477–8.

7. Mehta N, Allam S, Mazimba S, Karim S. Racial, ethnic, and socioeconomic disparities in out-of-hospital cardiac arrest within the United States: Now is the time for change. Heart Rhythm O2. 2022;16(3):857–63.

8. Mitchell WG, Deane A, Brown A, Bihari S, Wong H, Ramadoss R, et al. Long term outcomes for Aboriginal and Torres Strait Islander Australians after hospital intensive care. Medical Journal of Australia. 2020;213(1):16–21.

9. Greenwood BN, Carnahan S, Huang L. Patient–physician gender concordance and increased mortality among female heart attack patients. Proceedings of the National Academy of Sciences. 2018;115(34):8569–74.

10. Pantielieiev D. Garbage In, Garbage Out: How to Stop Your AI from Hallucinating Shelf.io2025 [Available from: <https://shelf.io/blog/garbage-in-garbage-out-ai-implementation/>.

11. Celi L, Cellini J, Charpignon M, Dee E, Dernoncourt F, Eber R, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities-A global review. PLOS Digital Health. 2022;1(3):e0000022.

12. Trout MJ, Henson G, Senthuran S. Characteristics and outcomes of critically ill Aboriginal and/or Torres Strait Islander patients in North Queensland. Medical Journal of Australia. 2016.

13. Stephens DP. Critical illness and its impact on the Aboriginal people of the top end of the Northern Territory, Australia. Journal of Anaesthesia and Intensive Care. 2003.

14. McDonald S, Russ G. Burden of end-stage renal disease among indigenous people in Australia and New Zealand. Kidney International. 2003;63:123–7.

15. Y Zhao CC, J Wright. Estimating Chronic Disease Prevalence Among the Remote Aboriginal Population of the Northern Territory. Australian New Zealand Journal of Public Health. 2008.

16. The health & welfare of Australia's Aboriginal & Torres Strait Islander people. <https://www.aihw.gov.au/reports-data/health-welfare-overview/indigenous-health-welfare/overview>: Australian Government; 2015.

17. Einsiedel L, Lersel Ev, Macnamara R, Spelman T, Heffernan M, Bray L, et al. Self-discharge by Aboriginal patients at Alice Springs Hospital, Central Australia: insights from a prospective cohort study. Australian Health Review. 2013;37(2):239–45.

18. Durey A, Thompson S, Wood M. Time to bring down the twin towers in poor Aboriginal hospital care: addressing institutional racism and misunderstandings in communication. Internal Medicine Journal. 2011;42(1):17–22.

19. Henry B, Houston S, Mooney G. Institutional racism in Australian healthcare: a plea for decency. Medical Journal of Australia. 2004;180(10).

20. Lee S, Ahn K, Cha M. Community-level socioeconomic status and outcomes of patients with out-of-hospital cardiac arrest

A systematic review and meta analysis. Medicine. 2021;3:e24170.

21. Nichol G. Regional Variation in Out-of-Hospital Cardiac Arrest Incidence and Outcome. JAMA. 2008;300(12):1423.

22. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53.

23. Teotia K, Jia Y, Woite N, Celi L, 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).

24. Jacobi J BN, Krinsley J, Agus M, Braithwaite S, Deutschman C, Freire A, Geehan D, Kohl B, Nasraway S, Rigby M, Sands K, Schallom L, Taylor B, Umpierrez G, Mazuski J, Schunemann H. Guidelines for the use of an insulin infusion for the management of hyperglycemia in critically ill patients. Crit Care Med. 2012;40(12):3251–76.

25. Krinsley JS, Bruns DE, Boyd JC. The Impact of Measurement Frequency on the Domains of Glycemic Control in the Critically Ill-A Monte Carlo Simulation. Journal of Diabetes Science and Technology. 2015;9(2):237–45.

26. R S, A M, G D, N A, S K, K R. Clinical challenges of glycemic control in the intensive care unit: A narrative review. World J Clin Cases. 2022;10(31):11260–72.

27. Falciglia M, Freyberg RW, Almenoff PL, D’Alessio DA, Render ML. Hyperglycemia–related mortality in critically ill patients varies with admission diagnosis\*. Critical Care Medicine. 2009;37(12):3001–9.

28. Oca S, Navas A, Leiman E, Buckland D. Effect of language interpretation modality on throughput and mortality for critical care patients: A retrospective observational study. J Am Coll Emerg Physicians Open 2021;2(4):e12477.

29. A D, D K, S C, T B, S T, D P. The impact of language barriers & interpreters on critical care patient outcomes. J Crit Care. 2023;73:154182.

30. Sekalala S, Chaiokobo T. Colonialism in the new digital health agenda. BMJ Global Health. 2024;9(e014131).

31. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Internal Medicine. 2018;178(11):1544.

32. Little R, Rubin D. Statistical Analysis with Missing Data, Third Edition. Wiley, editor2019.

33. Buuren Sv. Flexible imputation of missing data. Advanced guide on multiple and flexible imputation strategies. . 2012.

34. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5.

35. Abowd G, Dey A, Brown P, Davies N, Smith M, Steggles P. Towards a Better Understanding of Context and Context-Awareness. Handheld and Ubiquitous Computing HUC 1999 Lecture Notes in Computer Science; Berlin, Hiedelberg: Springer; 1999.

36. Acampora G, Cook DJ, Rashidi P, Vasilakos AV. A Survey on Ambient Intelligence in Healthcare. Proceedings of the IEEE. 2013;101(12):2470–94.

37. Lindahl C, Wagner S, Uldbjerg N, Schlütter JM, Bertelsen O, Sandager P. Effects of context-aware patient guidance on blood pressure self-measurement adherence levels. Health Informatics Journal. 2019;25(2):417–28.

38. Kim Y, Finn M. EPISTEMOLOGIES OF MISSING DATA: COVID DATA BUILDERS AND THE PRODUCTION AND MAINTENANCE OF MARGINALIZED COVID DATASETS. AoIR Selected Papers of Internet Research. 2023.

39. Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo L. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems: ACM Digital Library <https://dl.acm.org/doi/10.1145/3411764.3445518>; 2021. p. 1–15.

40. Oakden-Rayner L, Carneiro G, Dunnmon J, Ré C. Hidden Stratification Causes Clinically MeaningfulFailures in Machine Learning for Medical Imaging. Machine Learning for Health (ML4H) at NeurIPS 2019 2019.

41. Wang Y, Singh L. Analyzing the impact of missing values and selection bias on fairness. International Journal of Data Science and Analytics. 2021;12(2):101–19.

42. Dube L, Verster T. Assessing the Performance of Machine Learning Models for Default Prediction under Missing Data and Class Imbalance: A Simulation Study. ORION. 2024;40(1).

43. Getzen E, Ungar L, Mowery D, Jiang X, Long Q. Mining for equitable health: Assessing the impact of missing data in electronic health records. J Biomed Inform. 2023;139(104269).

44. Kiourtis A, Mavrogiorgou A, Kyriazis D, Themistocleous M. Acquiring the Ontological Representation of Healthcare Data Through Metamodeling Techniques. In: EMCIS, editor. EMCISLecture Notes in Business Information Processing: Springer; 2017.

45. Zhang B, Lemoine B, Mitchell M. Mitigating Unwanted Biases with Adversarial Learning. arXiv pre-print server. 2018.

46. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns. 2021;2(10):100347.

47. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al., editors. Model Cards for Model Reporting2019: ACM.

48. Silva PJ, Rahimzadeh V, Powell R, Husain J, Grossman S, Hansen A, et al. Health equity innovation in precision medicine: data stewardship and agency to expand representation in clinicogenomics. Health Research Policy and Systems. 2024;22(1).

49. Brown S, Saxena D, Wall P. Data collection in the global south: practical, methodological, and philosophical considerations. Information Technology for Development. 2024;30(3).

50. Shaw J, Sekalala S. Health data justice: building new norms for health data governance. npj Digital Medicine. 2023;6(1).

51. Rossi N, Golinelli L, Bersani F, Geraci M. A retrospective analysis of the factors associated with surgical checklist compliance using data from a local health unit in Italy, 2018–2021. Journal of Evaluation in Clinical Practice. 2023;29(8):1372–9.

52. Abdallah M, Hammad A, Staegemann D. A Data Collection Quality Model for Big Data Systems. In: IEEE, editor. 2023 International Conference on Information Technology (ICIT); Jordan: IEEE; 2023.

53. Wylie L, McConkey S. Insiders’ Insight: Discrimination against Indigenous Peoples through the Eyes of Health Care Professionals. Journal of Racial and Ethnic Health Disparities. 2019;6(1):37–45.

54. Leonard K, Russo S, Martinez A, McElroy L. Our CommonAgendaGlobal Digital Compact March 2023: CAREStatement for Indigenous Data Sovereignty. 2023.

55. McBride Kelly L, Wong D, Timothy A. Measuring what counts in Aboriginal and Torres Strait Islander care: a review of general practice datasets available for assessing chronic disease care. Australian Journal of Primary Health. 2024;30(4).

56. Geck MS, Cristians S, Berger-González M, Casu L, Heinrich M, Leonti M. Traditional Herbal Medicine in Mesoamerica: Toward Its Evidence Base for Improving Universal Health Coverage. Frontiers in Pharmacology. 2020;11.

57. Gray M, Williams K, Henderson R, Oster R, Samara W, Grant B, et al. Efforts to Introduce an Indigenous Identifier in a Canadian Provincial Health Authority. International Journal of Indigenous Health. 2024;20(1).

58. ONIX. Digital transformation made worry-free and reliable with automated data validation: ONIX; 2025 [Available from: <https://www.onixnet.com/blog/digital-transformation-made-worry-free-and-reliable-with-automated-data-validation/>.

59. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ. 2020:m3164.

60. Organisation WH. Module 1: Framework and metrics. 2017.

61. Ferryman K. The Dangers of Data Colonialism in Precision Public Health. Global Policy. 2021;12(S6):90–2.

62. Looi M-K. What should decolonisation of medical institutions look like? BMJ. 2023:p2257.

63. The Lancet Digital H. Decolonising health data. The Lancet Digital Health. 2023;5(8):e477.

64. Tolera A, Firdisa D, Roba HS, Motuma A, Kitesa M, Abaerei AA. Barriers to healthcare data quality and recommendations in public health facilities in Dire Dawa city administration, eastern Ethiopia: a qualitative study. Frontiers in Digital Health. 2024;6.

65. Davis A, Voelkel J, Remmers C, Adams J, McGlynn E. Comparing Kaiser Permanente Members to the General Population: Implications for Generalizability of Research. Perm J. 2023;27(2):87–98.

66. Umoh U, Nyoho E. A Fuzzy Intelligent Framework for Healthcare Diagnosis and Monitoring of Pregnancy Risk Factor in Women. Journal of Health, Medicine and Nursing. 2015;18.

67. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Medical Informatics and Decision Making. 2016;16(1).

68. Carroll S, Rodriguez-Lonebear D, Martinez A. Indigenous Data Governance: Strategies from United States Native Nations. Data Science Journal. 2019;18:31.

69. Cordes A, Bak M, Lyndon M, Hudson M, Fiske A, Celi LA, et al. Competing interests: digital health and indigenous data sovereignty. npj Digital Medicine. 2024;7(1).

70. Semarchy. Tackling Healthcare Data Quality Challenges: Issues, Risks, and Solutions Semarchy: Semarchy; 2025 [Available from: <https://semarchy.com/blog/tackling-healthcare-data-quality-challenges/>.

71. Bardhan S, Jenamani M, Routray A. Improving IoT Sensor Data Quality using Kalman Filter: The case of Cold Chain monitoring for Indian Mangoes. In: IEEE, editor. 7th International Conference for Convergence of Technology (I2CT): IEEE; 2022.

72. Antosz K, Knapčíková L, Husár J. Evaluation and Application of Machine Learning Techniques for Quality Improvement in Metal Product Manufacturing. Applied Sciences. 2024;14(22):10450.

73. Achieving Manufacturing Excellence with Data Observability acceldata: acceldata; 2024 [Available from: <https://www.acceldata.io/blog/achieving-manufacturing-excellence-with-data-observability>.

74. How to Deal With Missing Financial Data Quantpedia: Quantpedia; 2023 [Available from: <https://quantpedia.com/how-to-deal-with-missing-financial-data/>.

75. Prakash S. AI-Driven Cloud Solutions for Anti-Money Laundering (AML) Compliance with Graph Neural Networks and Behavioral Analytics. European-American Journals. 2025.

76. Okafor N, Ingle R, Okwudili Matthew U, Saunders M, Delaney DT. Assessing and Improving IoT Sensor Data Quality in Environmental Monitoring Networks: A Focus on Peatlands. IEEE Internet of Things Journal. 2024;11(24):40727–42.

77. Leonila T, Senthil GA, Geerthik S, Sowmiya R, Nithish J. Dynamic Water Quality Monitoring via IoT Sensor Networks and Machine Learning Technique. In: IEEE, editor. 2024 International Conference on Communication, Computing and Internet of Things (IC3IoT); India: IEEE; 2024.