frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 09 July 2025
DOI 10.3389/fmed.2025.1606254

Check for updates

OPEN ACCESS

EDITED BY
Penglin Ma,
Guiqian International General Hospital, China

REVIEWED BY
Zhongheng Zhang,
Sir Run Run Shaw Hospital, China
Zhou Feihu,
Chinese PLA General Hospital, China

*CORRESPONDENCE
Tristan Struja
✉ tstruja@mit.edu

†These authors share first authorship
‡These authors share last authorship

RECEIVED 04 April 2025
ACCEPTED 20 June 2025
PUBLISHED 09 July 2025

CITATION
Pradhan P, Haug FW, Abu Hussein NS, Moukheiber D, Moukheiber L, Moukheiber M, Moukheiber S, Weishaupt LL, Ellen JG, D'Couto H, Williams IC, Celi LA, Matos J and Struja T (2025) Potential source of bias in AI models: lactate measurement in the ICU in sepsis patients as a template.
Front. Med. 12:1606254.
doi: 10.3389/fmed.2025.1606254

COPYRIGHT
© 2025 Pradhan, Haug, Abu Hussein, Moukheiber, Moukheiber, Moukheiber, Moukheiber, Weishaupt, Ellen, D'Couto, Williams, Celi, Matos and Struja. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Potential source of bias in AI models: lactate measurement in the ICU in sepsis patients as a template

Pratiksha Pradhan[1,2†], Fredrik Willumsen Haug[1,3†], Nebal S. Abu Hussein[1,4†], Dana Moukheiber[1], Lama Moukheiber[1], Mira Moukheiber[5], Sulaiman Moukheiber[6], Luca Leon Weishaupt[1], Jacob G. Ellen[7], Helen D'Couto[8], Ishan C. Williams[9], Leo Anthony Celi [1,10,11], João Matos[1,12,13‡] and Tristan Struja [1,14,15*‡]

[1]Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, United States, [2]College of Engineering, Northeastern University, Boston, MA, United States, [3]Harvard John A. Paulson School of Engineering and Applied Sciences, Boston, MA, United States, [4]Pulmonary Critical Care Sleep Medicine Division, Yale School of Medicine, New Haven, CT, United States, [5]Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA, United States, [6]Department of Computer Science, Worcester Polytechnic Institute Computer Science, Worcester, MA, United States, [7]Harvard Medical School, Boston, MA, United States, [8]Division of Pulmonary, Critical Care, and Sleep Medicine, Georgetown University Hospital, Washington, DC, United States, [9]School of Nursing, University of Virginia, Charlottesville, VA, United States, [10]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States, [11]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States, [12]Faculty of Engineering, University of Porto (FEUP), Porto, Portugal, [13]Institute for Systems and Computer Engineering, Technology and Science (INESCTEC), Porto, Portugal, [14]Medical University Clinic, Kantonsspital Aarau, Aarau, Switzerland, [15]Hospital Muri, Muri Aargau, Switzerland

Objective: Health inequities may be driven by demographics such as sex, language proficiency, and race-ethnicity. These disparities may manifest through likelihood of testing, which in turn can bias artificial intelligence models. We aimed to evaluate variation in serum lactate measurements in the intensive care unit (ICU) in sepsis.

Methods: Utilizing MIMIC-IV (2008–2019), we identified adults fulfilling sepsis-3 criteria. Exclusion criteria were ICU stay <1-day, unknown race-ethnicity, <18 years of age, and recurrent ICU-stays. Employing targeted maximum likelihood estimation analysis, we assessed the likelihood of a lactate measurement on day 1. For patients with a measurement on day 1, we evaluated the predictors of subsequent readings.

Results: We studied 15,601 patients (19.5% racial-ethnic minority, 42.4% female, and 10.0% limited English proficiency). After adjusting for confounders, Black patients had a slightly higher likelihood of receiving a lactate measurement on day 1 [odds ratio 1.19, 95% confidence interval (CI) 1.06–1.34], but not the other minority groups. Subsequent frequency was similar across race-ethnicities, but women had a lower incidence rate ratio (IRR) 0.94 (95% CI 0.90–0.98). Patients with elective admission and private insurance also had a higher frequency of repeated serum lactate measurements (IRR 1.70, 95% CI 1.61–1.81 and 1.07, 95% CI, 1.02–1.12, respectively).

Conclusion: We found no disparities in the likelihood of a lactate measurement among patients with sepsis across demographics, except for a small increase for Black patients, and a reduced frequency for women. Subsequent analyses should account for the variation in biomarker monitoring being present in MIMIC-IV.

KEYWORDS

sepsis, lactate, MIMIC-IV, critical care, health equity

Introduction

Disparities in healthcare are widely recognized, especially regarding discrimination based on race and ethnicity (1, 2). Such disparities can unveil themselves as differences in quality of care, unequal medical device performance, or access to services reflecting structural inequities (3). These biases are not only harmful for patient care, but can also impact the development of machine learning-based clinical algorithms that train on electronic health records (EHR) (4).

Ensuring the development of fair AI models is crucial, and addressing missing information is a key initial step in achieving this objective, especially when such information is not missing at random (5, 6). Unfortunately, this variation in the level of monitoring is often not taken into consideration in the development of machine learning-based clinical algorithms. In a 2017 study that evaluated 107 electronic health record (EHR)-based risk prediction tools, 49 did not account for missing data (7). A common approach to imputation is the use of normal values based on the assumption that laboratory tests that are not ordered are presumed to be within normal range, a practice that likely introduces bias (8).

The probability of detecting an abnormal finding is contingent on the frequency of testing. Consequently, non-randomly missing data can lead to spurious correlations—non-causal relationships between features and outcome—that are learned and then incorporated into clinical algorithms (9). When the etiology of missing data stems from social determinants of care, these biases can become ingrained in subsequent AI models, perpetuating, and even scaling existing disparities (10, 11). This is even more important in a high-stake environment such as in patients with sepsis admitted to the intensive care unit (ICU).

Sepsis is a severe life-threatening systemic infection and effective management of this condition requires prompt diagnosis, aggressive treatment, and continuous monitoring. Despite current advances, one key challenge remains the timely delivery of care. Herein, serum lactate level is one of the two key diagnostic tools of septic shock according to the guidelines (12, 13). Disparities in sepsis outcomes are known to exist (14). However, the drivers of sepsis disparities are unknown and the question of whether disparities extend to serum lactate monitoring remains underexplored.

This paper seeks associations between race and ethnicity, sex, and language and the frequency of serum lactate determination during the management of sepsis in the ICU. By shedding light on this dimension of care, we aim to contribute to a more comprehensive understanding of the social patterning of the data generation process in healthcare.

Methods

This observational retrospective study is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (15). The health equity language, narrative, and concepts of this paper follows the American Medical Association's recommendations (16).

Data extraction

Data was extracted from the publicly available MIMIC-IV database (17). The MIMIC database is maintained by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology and shared via the PhysioNet platform (18). The dataset has been de-identified, and the institutional review boards of the Massachusetts Institute of Technology (No. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) both approved the use of the database for research. The MIMIC-IV database includes physiologic data collected from bedside monitors, laboratory test results, medications, medical images, and clinical progress notes captured in the electronic health record from patients admitted to the ICU between 2008 and 2019.

Hypothesis

We hypothesized that both the likelihood for a patient to have a serum lactate measurement and the frequency of subsequent measurements are not the same across race-ethnicity, sex, and English proficiency (as recorded by providers).

Cohort selection

The following exclusion criteria were applied to create a study cohort: those without sepsis as defined by the sepsis-3 criteria (12), patients under 18 years of age, and those with length of ICU stay [length of stay (LOS)] <1 day. Patients with recurrent ICU-stays in the database, and those with a racial description other than White, Asian, Black, or Hispanic, especially excluding those of the heterogenous group "other." For the negative binomial regression,

we further excluded patients with absent serum lactate values on day 1.

Covariates

We drew directed acyclic graphs (DAG) to understand which variables to extract (Supplementary Figure S1, Supplementary Table S1). Twelve confounders were extracted, including non-time-varying variables such as demographics, comorbidities, admission information, and source of infection and time-varying variables including Sequential Organ Failure Assessment (SOFA) score (19), and fluids normalized by LOS. Time-varying variables were modeled as follows: SOFA score was calculated for the day of ICU admission; serum lactate measurements were used as a binary variable for whether or not it was measured on day 1, in addition to taking the overall number of measurements for the whole ICU stay normalized by LOS.

Outcomes

We had two primary outcomes: the first was a binary variable predicting whether a patient received serum lactate measurement on day 1; the second was the number lactate measurements a patient would receive divided by the number of days in the ICU (LOS).

Statistical analysis

Statistical analysis was performed using Python 3.10.9 (20) and R 4.2.1 (21). For the outcome of whether or not a patient had a serum lactate measurement on day 1, we fitted a Targeted Maximum Likelihood Estimation (TMLE) model (22). From the TMLE model, we extracted and utilized the odds ratio (OR) to estimate the odds of receiving a serum lactate measurement. For the outcome of the number of serum lactate measurements during an ICU stay, we fitted a non-penalized, negative binomial regression [*statsmodel* package (23)] adjusted for confounders to estimate the number of serum lactate measurements for each patient each day in the ICU. We report our findings as incident rate ratios (IRR). All findings are reported with 95% CI and with White patients as the reference group.

Results

Baseline study cohort

The MIMIC-IV database has 73,140 ICU stays, of which 15,601 were included in our final cohort following application of the inclusion and exclusion criteria (Figure 1). The race-ethnicity distribution was 10.8% Black, 3.8% Hispanic, 2.9% Asian, 68.8% White and 14.6% others (without specified race). The demographic distribution did not change after applying exclusion criteria.

SOFA score had a median of 6.00 (interquartile range (IQR) 4.00, 8.00; Table 1), regardless of the race-ethnicity reported at baseline, with the Charlson comorbidity index at 6.00 (IQR 4.00, 8.00). Serum lactate on day 1 was slightly higher in the Non-White group at 2.50mmol/L (IQR 1.60, 4.00), compared to the White group at 2.20 (1.50, 3.50). In addition, Non-White patients received more fluids on the first day in the ICU than White patients [2,060 ml (IQR 640, 5,000) vs. 1,690 (461, 4,540)], respectively. Of note, the volume of fluids received prior to admission to the ICU is not available in the dataset. Upon a reviewer's request, we added a supplementary breakdown of baseline characteristics and illness severity by race-ethnicity (White, Black, and non-White/non-Black) to enhance transparency regarding racial representation in the dataset (Supplementary Table S2).

Model results

We adjusted our models for confounders according to a DAG (Supplementary Figure S1, Supplementary Table S1). Using the TMLE model with being White, male and English proficient as a reference, Black patients were more likely to have a serum lactate measurement on day 1 with OR 1.19 (95% CI 1.06, 1.34). Asian and Hispanic patients had a similar likelihood compared to White patients, with an OR of 1.08 (95% CI 0.93, 1.24), and an of OR 0.98 (95% CI 0.89, 1.08), respectively (Table 2, Figure 2a). We validated these findings with a cross-validated logistic regression model (Supplementary Table S3).

The negative binomial model was fitted to predict the total frequency of serum lactate measurements during a patient's ICU stay (Table 3, Figure 2b). We found no significant difference in the frequency of measurements across race-ethnicities compared to white patients as reference. Hispanic (IRR 1.12, 95% CI 0.99, 1.26), Black (IRR 1.01, 95% CI 0.94, 1.09), and Asian (IRR 1.08, 95% CI 0.95, 1.23) patients had a non-significant difference in their frequency of serum lactate measurements. In addition, English proficiency had no significant impact on measurement frequency (IRR 1.06, 95% CI 0.97, 1.16). On the other hand, female sex (IRR 0.94, 95% CI 0.90, 0.98) and having a urinary tract infection (IRR 0.68, 95% CI 0.50, 0.93) were associated with a decreased serum lactate measurement frequency, while having private insurance (IRR 1.07, 95% CI 1.02, 1.12) and being admitted electively (IRR 1.7, 95% CI 1.61, 1.81) significantly increased the frequency of receiving a measurement. Further, we conducted additional stratified analyses by admission SOFA score [cutoff at median of 6 supported by Ke et al. (24)] to explore differences in lactate measurement frequency by race and gender within illness severity strata (Supplementary Tables S4, S5).

Discussion

In this retrospective cohort study in patients with sepsis, we observed no discernible disparities between sexes and non-native English speakers in receiving a serum lactate measurement on day 1, although Black patients had a slightly increased likelihood. Furthermore, no apparent racial or language disparities were evident when examining the frequency of subsequent measurements, although a lower frequency was observed for women, those with private insurance, and those admitted electively.

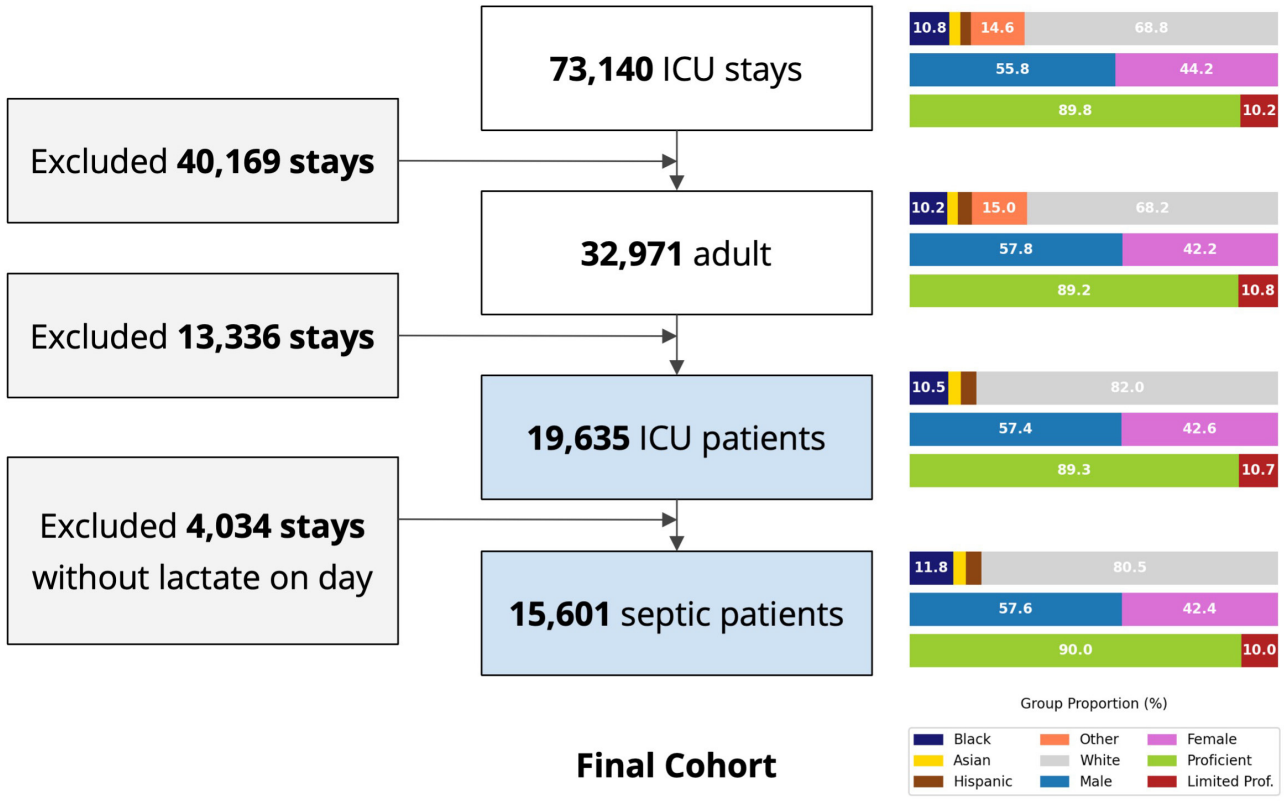

FIGURE 1
Study cohort selection flow chart, MIMIC-IV. Right panels depict the change of key demographic factors through application of the exclusion criteria. ICU, intensive care unit; LOS, length of stay; Proficient, English proficient; Limited Prof., limited English proficiency.

As Non-white patients were more likely to have Medicaid, there might still be disparities in care not captured in our data.

Although our study does not directly involve AI model development, its findings are highly relevant to the growing use of clinical data in artificial intelligence applications. Variability in measurement frequency, such as with lactate, can introduce biases into model training and deployment, particularly if the data reflects healthcare process differences rather than true physiological states. This is especially important given recent concerns about fairness and generalizability in AI models, which often underperform in underrepresented patient populations due to uneven data quality and representation (25, 26). Understanding and quantifying these real-world data characteristics is therefore, essential for building equitable and reliable AI systems in critical care. Health equity has become a priority in clinical research and among policymakers not only in the US but globally (27–29). In recent years, significant legislative changes around AI and health equity outcomes have been proposed and implemented. The European Parliamentary Research Service conducted a study on AI in healthcare in 2022 and recommended the implementation of specific coordination and support programs to address issues pertaining to AI and bias (30). In December 2023, the European Union approved the world's first legislation to regulate AI (31).

Beyond the obvious risks associated with feeding non-representative data to a model, variation in the clinical monitoring of patients presents a problem in the development of prediction, classification and optimization models using real-world data. The non-random sparsity of data from minoritized groups, even when represented in the dataset, has implications in the application of any statistical model. Machine learning-based decision support tools are an especially delicate area due to the sensitive nature of clinical decisions. Providers often intentionally refrain from measuring a variable especially in the ICU because of increasing recognition of the harm from over-testing (32). But the rationale behind such decisions is typically more complex, and confounded by both clinical and non-clinical (i.e., social determinants of care) features. In result, AI models learn wrong associations between clinical features and outcomes of interest. The problem becomes more pronounced in the advent of multi-modal modeling that requires black box deep learning representations (9). Models built on real-world data are thus subject to the human biases of the people who collected the primary data. For instance, a recent study found that large language models recommended low paying jobs more frequently to Mexicans, or implied that administrative work is solely a female job (33).

While our study did not assess mortality outcomes directly, partly due to the publication of similar work (34), the relationship between measurement frequency and patient outcomes remains an important area for future research. Causal inference methods, such as target trial emulation (35), could help determine whether

10.3389/fmed.2025.1606254

TABLE 1 Baseline information on the study cohort, derived from MIMIC-IV.

Variables	Missing	Race and ethnicity		
		Overall	Non-White	White
N (%)		15,601 (100)	2,801 (17.9)	12,800 (82.1)
Age, median [Q1, Q3]	0	68.0 (57.0, 78.0)	64.0 (52.0, 76.0)	68.0 (59.0, 79.0)
Female sex, *n* (%)	0	6,520 (41.8%)	1,341 (47.9%)	5,179 (40.5%)
English proficient, *n* (%)	0	14,113 (90.5%)	1,894 (67.6%)	12,219 (95.5%)
Insurance Medicaid, *n* (%)	0	1,042 (6.7%)	398 (14.2%)	644 (5.0%)
Insurance Medicare, *n* (%)	0	7,476 (47.9%)	1,064 (38.0%)	6,412 (50.1%)
Insurance Other, *n* (%)	0	7,083 (45.4%)	1,339 (47.8%)	5,744 (44.9%)
Charlson comorbidity index, median [Q1, Q3]	0	6.00 (4.00, 8.00)	6.00 (4.00, 8.00)	6.00 (4.00, 8.00)
SOFA, median [Q1, Q3]	0	6.00 (4.00, 8.00)	6.00 (4.00, 9.00)	6.00 (4.00, 8.00)
Elective admission, *n* (%)		2,876 (18.4%)	312 (11.1%)	2,564 (20.0%)
Length of stay, days, median [Q1, Q3]	0	3.13 (1.83, 6.25)	3.21 (1.88, 6.83)	3.13 (1.83, 6.17)
Lactate day 1 (mmol/L), median [Q1, Q3]	0	2.20 (1.50, 3.50)	2.50 (1.60, 4.00)	2.20 (1.50, 3.40)
Number of lactate measurements day 1, median [Q1, Q3]		3.00 (2.00, 5.00)	3.00 (2.00, 5.00)	3.00 (2.00, 5.00)
Lactate day 2 (mmol/L), median [Q1, Q3]	9,397 (60.2%)	1.70 (1.20, 2.60)	1.80 (1.30, 2.90)	1.70 (1.20, 2.60)
Number of lactate measurements day 2, median [Q1, Q3]	9,397 (60.2%)	2.00 (1.00, 3.00)	2.00 (1.00, 3.00)	2.00 (1.00, 3.00)
Mechanical ventilation, *n* (%)	0	8,841 (56.7%)	1,566 (55.9%)	7,275 (56.8%)
Renal replacement therapy, *n* (%)	0	1,550 (9.9%)	397 (14.2%)	1,153 (9.0%)
Vasopressor(s), *n* (%)	0	9,243 (59.2%)	1,455 (51.9%)	7,788 (60.8%)
Fluids received day 1 (ml), median [Q1, Q3]	446 (2.9%)	1,750 (498, 4,620)	2,060 (640, 5,000)	1,690 (461, 4,540)

Q1, lower quartile range; Q3, upper quartile range; SOFA, sequential organ failure assessment.

TABLE 2 Likelihood of receiving a lactate measurement on day 1 fitted by a targeted maximum likelihood estimation (TMLE) model.

Demographic	OR	2.50% CI	97.5% CI
White	Reference		
Black	1.19	1.06	1.34
Asian	1.08	0.93	1.24
Hispanic	0.98	0.89	1.08
Male	Reference		
Female	1.02	0.96	1.09
English proficient	Reference		
English non-proficient	0.96	0.86	1.07

OR, odds ratio; CI, confidence interval.

more frequent monitoring leads to improved outcomes. In an effort to mitigate biases, some studies have suggested the use of causal inference frameworks for machine learning (33, 36, 37), which should help understand and avoid embedding biases into AI algorithms. Evaluating data inputs used in AI models for biases and disparities as done in our work is a prerequisite even before employing causal inference frameworks and should become standard practice as the understanding gained aids in building better, more equitable, and trustworthy AI models. This study provides a framework and approach for future work, as health care professionals, engineers, and developers have the moral accountability to ensure safe deployment of AI models (38, 39).

Lactate is frequently measured as part of bundled panels, such as point-of-care arterial blood gas analyses leading to synchronous measurement with other parameters. However, it can also be assessed independently in the central laboratory. In a previous study on blood glucose monitoring in the ICU, we compared point-of-care and central lab measurements and found no clinically relevant differences between the two methods (40), suggesting that the impact of measurement context may be limited in this setting. While our analysis was conducted using data from the MIMIC database, similar findings have been reported in other critical care datasets, including eICU and Duke. For instance, Matos et al. (41) observed patterns of variability in arterial blood gas measurements across these cohorts. This cross-dataset consistency suggests that our conclusions may be generalizable to other ICU populations and settings. The optimal frequency of monitoring of serum lactate measurement is unknown. Several recent studies and reviews have shown that serial lactate measurements and trends—such as peak levels, area under the curve, and clearance are associated with mortality in sepsis. For example, a 2023 nationwide Korean cohort study found that combining serial lactate values with SOFA improved mortality prediction compared to SOFA alone (42). A meta-analysis with roughly 4,400 patients suggests that a protocol focusing on lactate clearance leads to lower in-hospital mortality compared to ScvO2 normalization or usual care (43).

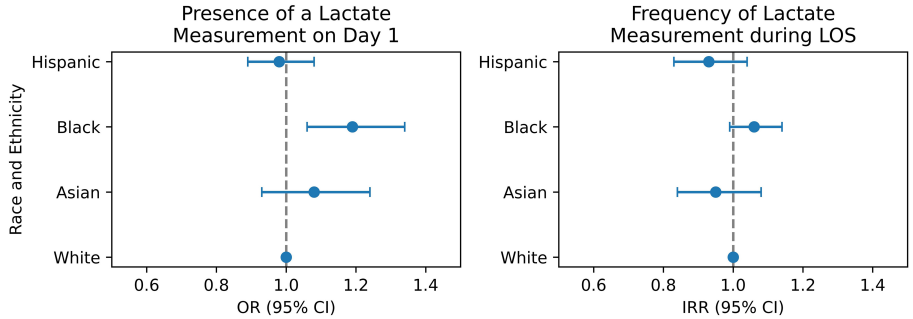

FIGURE 2
Summary of main results from TMLE **(left panel)** and negative binomial regression **(right panel)**. OR, odds ratio; IRR, incidence rate ratio; CI, confidence interval.

TABLE 3 Results of the negative binomial regression for outcome of lactate measurement frequency on day 1.

Variable	IRR	2.5% CI	97.5% CI
Intercept	0.72	0.62	0.85
Age	1.00	1.00	1.00
Charlson comorbidity index	1.01	1.00	1.02
SOFA	1.10	1.09	1.10
Volume of fluids normalized by LOS	1.00	1.00	1.00
Race			
White	Reference		
Asian	1.08	0.95	1.23
Black	1.01	0.94	1.09
Hispanic	1.12	0.99	1.26
Binary variables			
Female sex	0.94	0.90	0.98
English proficient	1.06	0.97	1.16
Private insurance	1.07	1.02	1.12
Elective admission	1.70	1.61	1.81
Pneumonia	1.01	0.90	1.13
Urinary tract infection	0.68	0.50	0.93
Biliary infection	1.22	0.81	1.84
Skin infection	1.03	0.61	1.72

IRR, incidence rate ratio; CI, confidence interval; SOFA, sequential organ failure assessment; LOS, length of stay.

A 2024 study in septic shock patients also linked initial, peak, and final 24-h lactate levels to 28-day mortality, although 24-h clearance was not predictive and optimal measurement frequency remained unclear (44). Similarly, a 2022 study using MIMIC-IV reported that lactate peak and AUC were associated with mortality, but lactate clearance was not more predictive than single values and performed worse than SOFA or NEWS scores (34). The current Surviving Sepsis Campaign guidelines support serial lactate monitoring but acknowledge that evidence for improving patient-centered outcomes remains limited and inconsistent (13). Many EHRs have already incorporated automated sepsis alerts to clinicians which rely on data such as the lactate to be present; disparities in collecting the data leads to disparities in usage of such alerts (45, 46). As such, the inputted data must be evaluated for bias. Other studies have shown that racially diverse Non-White ICU patients have nearly double the incidence of sepsis and higher rates of sepsis-related mortality compared to White patients (45, 47, 48). Furthermore, some studies in pediatric patients have reported higher mortality rates for those of lower socioeconomic status in the ICU (49). As such, all possible efforts need to be undertaken to close this disparity in patient care.

Limitations

While our research provides valuable insights into the discourse on disparities and biases within critical care, it is essential to acknowledge the limitations of our study. Firstly, selection bias could be a potential concern, as our data only encompassed patients admitted to the ICU in an academic tertiary care center in the USA whose patients are predominantly White. However, data from MIMIC is generally very similar to data from eICU-CRD another publicly available database encompassing 208 ICUs in the US (50). In general, race-ethnicity is self-reported in MIMIC-IV or provided by relatives, however in instances where this was not possible, data is recorded by the providers themselves. Additionally, our study design precludes us from testing for unmeasured confounding variables, especially leading to confounding by indication. Future research endeavors should make concerted efforts to address these limitations, such as including Social Determinants of Health and fostering a more comprehensive understanding of the topic by employing causal inference frameworks as the next prerequisite step before validating AI models. Although our observed variability in lactate measurements was statistically significant, the absolute magnitude was small, mostly due to the large sample size, and may not be clinically meaningful. However, even small differences can carry relevance in the critical care setting, where clinical decisions often hinge on marginal changes. Recent studies

have highlighted the challenges of bias and data completeness in electronic health records, particularly in underrepresented populations, which can impact both model performance and clinical interpretation (25, 26, 51). Our analysis focused on lactate due to its clinical relevance in sepsis and its frequent measurement in routine care, which allowed for robust assessment of intra-patient variability, but we agree that we cannot exclude incidental findings due to random variations. Also, similar disparities may affect other laboratory parameters (41), and our findings may not be generalizable to those. Future studies should explore a broader range of biomarkers to assess the extent of this issue across different clinical contexts. Moreover, future studies should extend their scope to cover other facets of care, including emergency departments, regular wards, or ambulatory care, to provide a more holistic perspective.

Conclusion

The implications of our study extend beyond the realm of lactate monitoring during sepsis management. In addition to the ongoing challenge of achieving healthcare equity within a system marked by systemic biases, clinicians and researchers must remain cognizant of these disparities before endeavoring to enhance patient care at their local institution or constructing any AI model. These biases not only have the potential to distort predictions, but may also endanger patient's safety when the predictions are employed for treatment or management decisions.

Data availability statement

The data that support the findings of this study are available in MIMIC-IV with the identifier doi.org/10.1093/jamia/ocx084 publicly available on PhysioNet (https://physionet.org). The code that produces the results in this manuscript can be accessed at https://github.com/joamats/mit-lactate, which includes detailed instructions for running the code.

Ethics statement

The studies involving humans were approved by Massachusetts Institute of Technology (No. 0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14). The studies were conducted in accordance with the local legislation and institutional requirements. The Ethics Committee/Institutional Review Board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because of the retrospective nature with no change of treatment.

Author contributions

PP: Formal analysis, Data curation, Conceptualization, Writing – review & editing, Investigation. FW: Writing – review & editing, Formal analysis, Data curation. NAH: Writing – original draft, Writing – review & editing, Validation. DM: Data curation, Conceptualization, Writing – review & editing. LM: Conceptualization, Writing – review & editing, Data curation. MM: Writing – review & editing, Conceptualization, Data curation. SM: Data curation, Writing – review & editing. LW: Data curation, Conceptualization, Writing – review & editing, Formal analysis. JE: Writing – review & editing, Visualization. HD'C: Supervision, Writing – review & editing. IW: Writing – review & editing. LC: Conceptualization, Writing – review & editing. JM: Supervision, Formal analysis, Writing – review & editing, Data curation, Conceptualization. TS: Supervision, Data curation, Conceptualization, Writing – review & editing, Formal analysis, Validation, Funding acquisition, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. LAC 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. JM was supported by a Fulbright/FLAD Grant, Portugal, AY 2022/2023. TS is supported by the Swiss National Science Foundation (P400PM_194497). NAH is supported by the Swiss National Science Foundation (P500PM_210847).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2025.1606254/full#supplementary-material

References

1. Magesh S, John D, Li WT, Li Y, Mattingly-App A, Jain S, et al. Disparities in COVID-19 outcomes by race, ethnicity, and socioeconomic status: a systematic review and meta-analysis. *JAMA Netw Open*. (2021) 4:e2134147. doi: 10.1001/jamanetworkopen.2021.34147

2. Hall WJ, Chapman MV, Lee KM, Merino YM, Thomas TW, Payne BK, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *Am J Public Health*. (2015) 105:e60–76. doi: 10.2105/AJPH.2015.302903

3. Charpignon ML, Byers J, Cabral S, Celi LA, Fernandes C, Gallifant J, et al. Critical bias in critical care devices. *Crit Care Clin*. (2023) 39:795–813. doi: 10.1016/j.ccc.2023.02.005

4. Ferryman K, Mackintosh M, Ghassemi M. Considering biased data as informative artifacts in AI-assisted health care. *N Engl J Med*. (2023) 389:833–8. doi: 10.1056/NEJMra2214964

5. Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health*. (2023) 2:e0000278. doi: 10.1371/journal.pdig.0000278

6. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci*. (2021) 4:123–44. doi: 10.1146/annurev-biodatasci-092820-114757

7. Gold R, Cottrell E, Bunce A, Middendorf M, Hollombe C, Cowburn S, et al. Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. *J Am Board Fam Med*. (2017) 30:428–47. doi: 10.3122/jabfm.2017.04.170046

8. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMs*. (2013) 1:1035. doi: 10.13063/2327-9214.1035

9. Yang Y, Zhang H, Katabi D, Ghassemi M. Change is hard: a closer look at subpopulation shift [Internet]. *arXiv*. (2023). Available online at: http://arxiv.org/abs/2302.12254 (Accessed October 15, 2024].

10. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. (2019) 9:010318. doi: 10.7189/jogh.09.020318

11. Martin K. *Ethics of Data and Analytics: Concepts and Cases*. 1st ed. Boca Raton, FL: CRC Press (2022). p. 1. doi: 10.1201/9781003278290

12. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*. (2016) 315:801. doi: 10.1001/jama.2016.0287

13. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Crit Care Med*. (2021) 49:e1063–143. doi: 10.1097/CCM.0000000000005337

14. Black LP, Hopson C, Puskarich MA, Modave F, Booker SQ, DeVos E, et al. Racial disparities in septic shock mortality: a retrospective cohort study. *Lancet Reg Health*. (2024) 29:100646. doi: 10.1016/j.lana.2023.100646

15. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. (2008) 61:344–9. doi: 10.1016/j.jclinepi.2007.11.008

16. Flanagin A, Frey T, Christiansen SL, AMA Manual of Style Committee. Updated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA*. (2021) 326:621. doi: 10.1001/jama.2021.13304

17. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. (2023) 10:1. doi: 10.1038/s41597-023-01945-2

18. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. (2000) 101:E215–20. doi: 10.1161/01.CIR.101.23.e215

19. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure: on behalf of the working group on sepsis-related problems of the European Society of Intensive Care Medicine (see contributors to the project in the appendix). *Intensive Care Med*. (1996) 22:707–10. doi: 10.1007/BF01709751

20. van Rossum G. *Python reference manual*. CWI (1995).

21. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team (2022). Available from: https://www.R-project.org/

22. Gruber S, Laan MJVD. TMLE : an *R* package for targeted maximum likelihood estimation. *J Stat Softw*. (2012) 51:1–35. doi: 10.18637/jss.v051.i13

23. Seabold S, Perktold J. *Statsmodels: Econometric and Statistical Modeling with Python*. Austin, TX (2010) p. 92–6. doi: 10.25080/Majora-92bf1922-011

24. Ke Y, Tang MSS, Loh CJL, Abdullah HR, Shannon NB. Cluster trajectory of SOFA score in predicting mortality in sepsis. *arXiv*. (2023) 1–26. Available online at: http://arxiv.org/abs/2311.17066 (Accessed May 29, 2025).

25. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol*. (2024) 42:3–15. doi: 10.1007/s11604-023-01474-3

26. Allareddy V, Oubaidin M, Rampa S, Venugopalan SR, Elnagar MH, Yadav S, et al. Call for algorithmic fairness to mitigate amplification of racial biases in artificial intelligence models used in orthodontics and craniofacial health. *Orthod Craniofac Res*. (2023) 26:124–30. doi: 10.1111/ocr.12721

27. Centers for Medicare & Medicaid Services. *CMS Proposes Policies to Improve Patient Safety and Promote Health Equity* (2023).

28. Yao Q, Li X, Luo F, Yang L, Liu C, Sun J. The historical roots and seminal research on health equity: a referenced publication year spectroscopy (RPYS) analysis. *Int J Equity Health*. (2019) 18:152. doi: 10.1186/s12939-019-1058-3

29. Coley RY, Duan KI, Hoopes AJ, Lapham GT, Liljenquist K, Marcotte LM, et al. A call to integrate health equity into learning health system research training. *Learn Health Syst*. (2022) 6:e10330. doi: 10.1002/lrh2.10330

30. Lekadir K, Quaglio G, Garmendia AT, Gallin C. Artificial intelligence in healthcare-applications, risks, and ethical and societal impacts. *European Parliament* (2022). Available from: https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU(2022)729512_EN.pdf

31. European Parliament. *EU AI Act: First Regulation on Artificial Intelligence | News |* (2023). Available online at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (Accessed December 9, 2023).

32. Kox M, Pickkers P. "Less Is More" in critically ill patients: not too intensive. *JAMA Intern Med*. (2013) 173:1369. doi: 10.1001/jamainternmed.2013.6702

33. Scholkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. *Proc IEEE*. (2021) 109:612–34. doi: 10.1109/JPROC.2021.3058954

34. Wang M, Wang Y, Taotao L, Zhao Q, Chao Y. Evaluation of plasma lactate parameters for predicting mortality of septic patients. *Heliyon*. (2022) 8:e12340. doi: 10.1016/j.heliyon.2022.e12340

35. Yang J, Wang L, Chen L, Zhou P, Yang S, Shen H, et al. A comprehensive step-by-step approach for the implementation of target trial emulation: evaluating fluid resuscitation strategies in post-laparoscopic septic shock as an example. *Laparosc Endosc Robot Surg*. (2025) 8:28–44. doi: 10.1016/j.lers.2025.01.001

36. Struja T, Matos J, Lam B, Cao Y, Liu X, Jia Y, et al. Evaluating equitable care in the ICU: Creating a causal inference framework to assess the impact of life-sustaining interventions across racial and ethnic groups. *medRxiv*. [Preprint]. (2023) 2023:10.12.23296933. doi: 10.1101/2023.10.12.23296933

37. Plecko D, Bareinboim E. Causal fairness analysis. *arXiv*. (2022). Available online at: https://arxiv.org/abs/2207.11385 (Accessed October 15, 2024).

38. Xie Y, Zhuang D, Chen H, Zou S, Chen W, Chen Y. 28-day sepsis mortality prediction model from combined serial interleukin-6, lactate, and procalcitonin measurements: a retrospective cohort study. *Eur J Clin Microbiol Infect Dis*. (2023) 42:77–85. doi: 10.1007/s10096-022-04517-1

39. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. *Bull World Health Organ*. (2020) 98:251–6. doi: 10.2471/BLT.19.237487

40. 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. doi: 10.1016/j.jbi.2024.104643

41. Matos J, Alwakeel M, Hao S, Naamani D, Struja T, Gichoya JW, et al. Differences in arterial blood gas testing by race and sex across 161 US hospitals in 4 EHR databases. *Am J Respir Crit Care Med*. (2025) 211:1049–58. doi: 10.1164/rccm.202406-1242OC

42. Park H, Lee J, Oh DK, Park MH, Lim CM, Lee SM, et al. Serial evaluation of the serum lactate level with the SOFA score to predict mortality in patients with sepsis. *Sci Rep*. (2023) 13:6351. doi: 10.1038/s41598-023-33227-7

43. Simpson SQ, Gaines M, Hussein Y, Badgett RG. Early goal-directed therapy for severe sepsis and septic shock: a living systematic review. *J Crit Care*. (2016) 36:43–8. doi: 10.1016/j.jcrc.2016.06.017

44. Stoiber A, Hermann A, Wanka ST, Heinz G, Speidl WS, Hengstenberg C, et al. Enhancing SAPS-3 predictive accuracy with initial, peak, and last lactate measurements in septic shock. *J Clin Med*. (2024) 13:3505. doi: 10.3390/jcm13123505

45. Barnato AE, Alexander SL, Linde-Zwirble WT, Angus DC. Racial variation in the incidence, care, and outcomes of severe sepsis: analysis of population,

patient, and hospital characteristics. *Am J Respir Crit Care Med*. (2008) 177:279–84. doi: 10.1164/rccm.200703-480OC

46. Raman J, Johnson TJ, Hayes K, Balamuth F. Racial differences in sepsis recognition in the emergency department. *Pediatrics*. (2019) 144:e20190348. doi: 10.1542/peds.2019-0348

47. Mayr FB. Infection rate and acute organ dysfunction risk as explanations for racial differences in severe sepsis. *JAMA*. (2010) 303:2495. doi: 10.1001/jama.2010.851

48. Martin GS, Mannino DM, Eaton S, Moss M. The epidemiology of sepsis in the United States from 1979 through 2000. *N Engl J Med*. (2003) 348:1546–54. doi: 10.1056/NEJMoa022139

49. Reddy AR, Badolato GM, Chamberlain JM, Goyal MK. Disparities associated with sepsis mortality in critically ill children. *J Pediatr Intensive Care*. (2022) 11:147–52. doi: 10.1055/s-0040-1721730

50. Sauer CM, Dam TA, Celi LA, Faltys M. de la Hoz MAA, Adhikari L, et al. Systematic review and comparison of publicly available ICU data sets-a decision guide for clinicians and data scientists. *Crit Care Med*. (2022) 50:e581–8. doi: 10.1097/CCM.0000000000005517

51. Ellen JG, Matos J, Viola M, Gallifant J, Quion J, Anthony Celi L, et al. Participant flow diagrams for health equity in AI. *J Biomed Inform*. (2024) 152:104631. doi: 10.1016/j.jbi.2024.104631