Raising awareness of potential biases in medical machine learning: Experience from a Datathon

Harry Hochheiser[1*], Jesse Klug[2], Thomas Mathie[3], Tom J. Pollard[4], Jesse D. Raffa[4], Stephanie L. Ballard[5], Evamarie A. Conrad[5], Smitha Edakalavan[1], Allan Joseph[6,7], Nader Alnomasy[5,8] Sarah Nutman[3], Veronika Hill[5], Sumit Kapoor[3], Eddie Pérez Claudio[1], Olga V. Kravchenko[9], Ruoting Li[3], Mehdi Nourelahi[1], Jenny Diaz[5], W. Michael Taylor[3], Sydney R. Rooney[10], Maeve Woeltje[3], Leo Anthony Celi[4,11,12], Christopher M. Horvat[3]

1 Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA USA
2 UPMC Intensive Care Unit Service Center, UPMC, Pittsburgh, PA, USA
3 Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, USA
4 MIT Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA
5 Health Informatics, School of Health and Rehabilitation Sciences, University of Pittsburgh, Pittsburgh, PA, USA
6 Division of Critical Care Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
7 Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA
8 College of Nursing, Medical Surgical Department, University of Hail, Ha'il, Saudi Arabia
9 Department of Family and Community Medicine, University of Pittsburgh, Pittsburgh, PA, USA
10 Division of Cardiology, Department of Pediatrics, Children's Hospital of Pittsburgh, University of Pittsburgh, Pittsburgh, PA, USA
11 Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
12 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

* harryh@pitt.edu

Abstract

Objective: To challenge clinicians and informaticians to learn about potential sources of bias in medical machine learning models through investigation of data and predictions from an open-source severity of illness score.
Methods: Over a two-day period (total elapsed time approximately 28 hours), we conducted a datathon that challenged interdisciplinary teams to investigate potential sources of bias in the Global Open Source Severity of Illness Score. Teams were invited to develop hypotheses, to use tools of their choosing to identify potential sources of bias, and to provide a final report.
Results: Five teams participated, three of which included both informaticians and clinicians. Most (4/5) used Python for analyses, the remaining team used R. Common analysis themes included relationship of the GOSSIS-1 prediction score with demographics and care related variables; relationships between demographics and

outcomes; calibration and factors related to the context of care; and the impact of missingness. Representativeness of the population, differences in calibration and model performance among groups, and differences in performance across hospital settings were identified as possible sources of bias.
Discussion: Datathons are a promising approach for challenging developers and users to explore questions relating to unrecognized biases in medical machine learning algorithms.

Author summary

Disadvantaged groups are at risk of being adversely impacted by biased medical machine learning models. To avoid these undesirable outcomes, developers and users must understand the challenges involved in identifying potential biases. We conducted a datathon aimed at challenging a diverse group of participants to explore an open-source patient severity model for potential biases. Five groups of clinicians and informaticians used tools of their choosing to evaluate possible sources of biases, applying a range of analytic techniques and exploring multiple features. By engaging diverse participants with hands-on data experience with meaningful data, datathons have the potential to raise awareness of potential biases and promote best practices in developing fair and equitable medical machine learning models.

Introduction

Increased awareness is arguably the first, and most important, step toward reduction of undesirable biases in medical machine learning (ML). Although reporting tools [1, 2], checklists [3] and bias exploration libraries [4–6] provide some assistance, the management of bias in medical ML projects is still a largely manual task, requiring exploration of data and testing of specified hypotheses. Effective efforts will likely rely on multiple perspectives, particularly bridging gaps between clinicians with domain knowledge necessary for generation of hypothetical sources of bias and analysts with statistical and programming skills necessary to test those hypotheses. To explore approaches for raising awareness of these issues, we conducted and participated in a datathon that challenged groups of informaticians and clinicians to generate hypotheses regarding potential biases of the Global Open Source Severity of Illness Score (GOSSIS-1) prediction algorithm [7] and conduct analyses to explore those hypotheses.

Materials and methods

GOSSIS-1 was designed as a modern replacement for earlier intensive-care severity of illness scores. Noting problems in generalizability and performance degradation over time, the GOSSIS-1 team used datasets from the eICU database of over 200,000 ICU admissions in the US [8] and a comparable database of more than 60,000 ICU admissions in Australia and New Zealand [9] to develop a new model and demonstrate a proposed methodology for developing globally-applicable severity scores. Although logistic regression models based on demographics, labs, vitals, and APACHE score predictions led to models with good performance and calibration [7], potential biases in the data have not yet been thoroughly studied.

The datathon was intended primarily to challenge emerging researchers to consider questions relating to bias in predictive models. Participants (and co-authors) included medical residents and fellows, students, faculty, and staff from the University of Pittsburgh and UPMC. The datathon was held virtually, over 1.5 working days in

February 2024. The first day included background presentations on the GOSSIS model and data, an introduction to bias in machine learning, a description of the Prediction model Risk Of Bias ASsessment Tool (PROBAST) bias reporting framework [1], and a description of the datathon's goals and judging criteria. From 1pm onwards, participants met and conducted analyses. The second day of the datathon started with one hour of short talks about ML/AI in medicine, followed by three hours of open time for working on assignments.

The teams had three options for accessing data to be used in the Datathon. Two datasets from PhysioNet were available: 1) the Women in Data Science Datathon 2020 dataset [10, 11] was available both in raw form and in a version that could be analyzed using ChatGPT Advanced Data Analysis Features, and 2) those who had completed appropriate Physionet credentialing were eligible to use the eICU-CRD subset of the GOSSIS-1 dataset [11, 12], containing the US data used in developing the GOSSIS-1 score. A third dataset was a synthetic derivative of data used to implement the GOSSIS-1 score at UPMC. Participants who chose either the eICU-CRD dataset or the synthetic UPMC data set were not able to use the ChatGPT function directly with the data, due to restrictions on data sharing, but could still use generative AI tools to assist with code writing.

Participating teams were allowed to use tools of their choice for analyzing data and presenting results. They were asked to develop hypotheses regarding potential sources of bias, using PROBAST questions regarding predictors, outcomes, and analysis as a framework for guiding hypothesis development. A grading rubric was provided and teams were instructed to provide a report detailing their hypotheses, features used, indications of any supplemental data, descriptions of analyses conducted, presentation of results, and descriptions of conclusions. Submissions were analyzed to extract tools used and to categorize analysis approaches used and resulting conclusions into common themes. Datathon materials, including introductory presentations, instructions, and submissions from each of the five teams can be found at https://doi.org/10.5281/zenodo.13962037.

Results

Sixteen participants were assigned into five teams of equal size and comparable distributions of backgrounds. As several registered participants either did not participate at all or did not continue to completion, final team sizes ranged from 1 to 5 members. Three of the teams included both clinicians and informaticians. One of the remaining teams had two members, both clinicians; the other team included one participant, an informatician. Slightly over one-half (9/16) of the participants had a clinical background. Detailed information on participants' experience with coding or data analytics was not collected.

Of the five teams, two used the synthetic UPMC dataset, two used the women in science dataset, and one team used both. No teams used the eICU-CRD dataset. Python was the preferred analysis tool, selected by four of the five teams. The remaining team used R. At least one team used a GUI-based tool for exploratory data analysis, and one team acknowledged using ChatGPT to assist with coding. Most (4) of the teams used Jupyter notebooks, one team used R.Four of the teams formatted their final report as a PDF document; the remaining team created a slide presentation. Teams used a variety of approaches for data presentation, including bar graphs, AUROC/AUPRC graphs, calibration plots, and heatmaps.

Participating teams used a variety of approaches to examine potential sources of bias. Three of the five teams provided explicitly stated hypotheses; four of the teams referred to PROBAST questions either in the formulation of hypotheses or in the presentation of

Table 1. Data features examined for potential biases

Feature/characteristic	# of teams (n=5)
GOSSIS predicted score	5
Age	4
Race/ethnicity	4
Mortality	4
Type of ICU/Hospital	3
Missingness	3
APACHE score	3
Calibration	3
Sex/gender	2
Comorbidities	1
Length of pre-ICU hospitalization	1
Data entry errors	1
Ventilation status	1
Feature importance	1
Readmission	1
Impact of imputation	1
ICU Admissions source	1

their conclusions. Age, race/ethnicity, and mortality were the most frequently examined features, considered by four of the five teams. A complete list of features or data characteristics is given in table 1.

Analysis approaches fell into three broad categories (Table 2). All 5 groups explored potential biases in GOSSIS scores including analyses of the relationships between predictions and demographics, context of care (hospital/ICU types), admission source, and exploration of calibration of GOSSIS predictions across ethnic groups. All groups also explored the impact of demographics (race, ethnicity, sex/gender, and age), including associations with outcomes, hospital type, length of pre-icu hospital stay, feature importance, and calibration, along with comparison of the patient population relative to the general population (as indicated by census data).

Teams used a variety of graphical and tabular approaches to present results. One team used an external data source (US Census Bureau information). Only one of the teams used any inferential statistical tests in their analyses. Although none of the teams found clear instances of bias, several areas of concern were identified, including possible biases in missingness of the data (n=2), representativeness of the population (2) , differences in calibration (3) and model performance among groups (3), and differences in performance across hospital settings (2).

Discussion

Despite significant awareness of the potential challenges and harms associated with inappropriately-biased medical AI/MLmodels, appropriate strategies for identifying and addressing bias are not as well understood. A range of techniques have been proposed to identify and ameliorate bias [13, 14] leading to the development of software libraries for measuring potential biases [4, 6] and decision tools designed to help researchers choose the most suitable fairness metrics for a given situation [5]. Several broad frameworks for addressing bias have been proposed, with varying levels of specificity [3, 15–17].Given the numerous possible sources of interactions, the complexity of biases they may introduce, and the need for additional data to provide context,

Table 2. Analysis Topics

Analysis Category	Subcategory	# of teams (n=5)
GOSSIS Score	Relationship with admission source (1), demographics (2), hospital/ICU type (2), admission source(1); association of calibration with ethnicity (3)	5
Demographics	Relationship with GOSSIS score (1), death/survival (1); Association with comorbidities (1), hospital type (1), length of pre-icu hospital stay (1), ventilation use (1), feature importance (1), and calibration (3) ; general distribution patterns (1)	5
Missingness	Association with demographics (1), GOSSIS score (1), hospital/ICU type (1); Impact of imputation on model performance (1)	3

identification of potential biases in machine learning models will likely remain an ongoing challenge requiring thoughtful examination of the models and the data. Educating ML developers and clinicians about the potential dangers of biased models is perhaps the most promising means of addressing this challenging problem.

Our datathon provides a model for engaging informaticians and clinicians in collaborative efforts to explore potential sources of inappropriately biased results from a machine learning model. Mixed groups involving participants with a range of clinical and technical experience were able to come together to quickly begin exploration of multiple hypotheses. Although the five teams identified different questions and used varying approaches to address those questions, interactions were collaborative and productive. The open-ended nature of the tasks and the straightforward data involved likely helped informaticians to generate relevant questions, despite their lack of clinical expertise. Similarly, flexibility of tools enabled clinicians who might not be well-versed in programming or statistical applications to use more familiar tools (although some clinicians were proficient R/Python programmers). Anecdotal observation of the groups in action suggested that exchanges between clinicians and informaticians helped participants better appreciate the complexities of the questions.

Teams generally shared a common focus on topics considered in their analyses and the features used to conduct those analyses. The relative lack of inclusion of statistical tests and external data sets (1 team for each) suggests that discussion of rigorous methods for quantifying the likelihood of bias and the potential utility of including additional data sources for contextualizing potential sources of bias might be useful topics for related training. Most (four of five) teams used the PROBAST model to frame their analyses. Similar detailed questions designed to guide analyses, or decision tools such as the fairness tree [5] might be helpful for guiding participants in future datathons, and for generally helping developers address questions of potential bias. None of the teams identified strong evidence for troubling bias. Given the limited available time, this is not surprising. However, they identified potentially concerning areas worthy of further consideration, representativeness of data, calibration, model performance, and the potential impact of context of care.

Beyond the framework provided by the PROBAST questions [1], participants were

given very little structure or guidance in developing hypotheses and exploring data. The open-ended nature of the task reflects the lack of structured approaches for conducting these analyses, and may have encouraged some creativity in addressing this problem. Future datathons might use a modified approach providing additional scaffolding in the form of structured questions, potentially encouraging the use of more formalized processes for identifying and testing hypotheses regarding potential biases.

Conclusion

Identification and reduction of inappropriate biases of medical machine learning models are increasingly important goals, likely requiring collaboration between developers and clinical users of those models. Our February 2024 datathon provides a model for using team-based investigation of meaningful datasets to explore hypotheses and identify potential sources of bias in need of further consideration. We hope to use subsequent datathons to explore possible systematic approaches for identifying biases in both underlying data and models trained using those data and improve methods to mitigate these biases to achieve equitable outcomes for all patients.

Acknowledgments

CMH and HH are supported by NIH Grant 5R01NS118716-03. CMH is also supported by 5K23HD099331. TP and LAC are supported by Nationals Institute of Health Grant OT2OD032701. LAC is also funded by NIH Grants R01 EB017205 and DS-I Africa U54 TW012043-01 and the National Science Foundation through ITEST #2148451. EPC is supported by NLM Training Grant T15LM007059.

References

1. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Annals of Internal Medicine. 2019;170(1):W1–W33. doi:10.7326/M18-1377.

2. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

3. Al-Zaiti SS, Alghwiri AA, Hu X, Clermont G, Peace A, Macfarlane P, et al. A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML). European Heart Journal - Digital Health. 2022;3(2):125–140. doi:10.1093/ehjdh/ztac016.

4. Weerts H, Dudík M, Edgar R, Jalali A, Lutz R, Madaio M. Fairlearn: Assessing and Improving Fairness of AI Systems. Journal of Machine Learning Research. 2023;24(257):1–8.

5. Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, et al.. Aequitas: A Bias and Fairness Audit Toolkit; 2019. Available from: `http://arxiv.org/abs/1811.05577`.

6. Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al.. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias; 2018. Available from: https://arxiv.org/abs/1810.01943v1.

7. Raffa JD, Johnson AEW, O'Brien Z, Pollard TJ, Mark RG, Celi LA, et al. The Global Open Source Severity of Illness Score (GOSSIS)*. Critical Care Medicine. 2022;50(7):1040. doi:10.1097/CCM.0000000000005518.

8. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data. 2018;5(1):180178. doi:10.1038/sdata.2018.178.

9. Stow PJ, Hart GK, Higlett T, George C, Herkes R, McWilliam D, et al. Development and implementation of a high-quality clinical database: the Australian and New Zealand Intensive Care Society Adult Patient Database. Journal of Critical Care. 2006;21(2):133–141. doi:10.1016/j.jcrc.2005.11.010.

10. Lee M, Raffa J, Ghassemi M, Pollard T, Kalanidhi S, Badawi O, et al.. WiDS (Women in Data Science) Datathon 2020: ICU Mortality Prediction; 2020. Available from: https://physionet.org/content/widsdatathon2020/.

11. Raffa J, Johnson A, Pollard T, Badawi O. GOSSIS-1-eICU, the eICU-CRD subset of the Global Open Source Severity of Illness Score (GOSSIS-1) dataset; 2022. Available from: https://physionet.org/content/gossis-1-eicu/1.0.0/.

12. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. 2000;101(23):e215–e220. doi:10.1161/01.CIR.101.23.e215.

13. Chen F, Wang L, Hong J, Jiang J, Zhou L. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. Journal of the American Medical Informatics Association. 2024;31(5):1172–1183. doi:10.1093/jamia/ocae060.

14. Ferrara E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci. 2023;6(1):3. doi:10.3390/sci6010003.

15. Abràmoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z, et al. Considerations for addressing bias in artificial intelligence for health equity. npj Digital Medicine. 2023;6(1):1–7. doi:10.1038/s41746-023-00913-9.

16. Chin MH, Afsar-Manesh N, Bierman AS, Chang C, Colón-Rodríguez CJ, Dullabh P, et al. Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. JAMA Network Open. 2023;6(12):e2345050. doi:10.1001/jamanetworkopen.2023.45050.

17. Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLoS Digital Health. 2023;2(6):e0000278. doi:10.1371/journal.pdig.0000278.