ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/381611973

The 2024 Pediatric Sepsis Challenge: Predicting In-Hospital Mortality in Children With Suspected Sepsis in Uganda

Article *in* Pediatric Critical Care Medicine · June 2024

DOI: 10.1097/PCC.0000000000003556

CITATIONS
0

READS
58

13 authors, including:

Charly Huxford
University of British Columbia
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Alireza Rafiei
Emory University
15 PUBLICATIONS **136** CITATIONS
SEE PROFILE

Matthew O Wiens
University of British Columbia
146 PUBLICATIONS **3,794** CITATIONS
SEE PROFILE

Mark Ansermino
University of British Columbia
512 PUBLICATIONS **9,870** CITATIONS
SEE PROFILE

All content following this page was uploaded by Mark Ansermino on 13 August 2024.

The user has requested enhancement of the downloaded file.

The 2024 Pediatric Sepsis Challenge: Predicting In-Hospital Mortality in Children With Suspected Sepsis in Uganda

Charly Huxford, BA[1]
Alireza Rafiei, MS[2]
Vuong Nguyen, PhD[1]
Matthew O. Wiens, PhD[1,3,4]
J. Mark Ansermino, MBBCh[1,3,4]
Niranjan Kissoon, MD[1,4,5]
Elias Kumbakumba, MMed[6]
Stephen Businge, MMed[7]
Clare Komugisha, BSc[8]
Mellon Tayebwa, MA[8]
Jerome Kabakyenga, PhD[9,10]
Nathan Kenya Mugisha, MMed[8]
Rishikesan Kamaleswaran, PhD[11,12]
on behalf of the Pediatric Sepsis Data CoLaboratory

ABSTRACT: The aim of this "Technical Note" is to inform the pediatric critical care data research community about the "2024 Pediatric Sepsis Data Challenge." This competition aims to facilitate the development of open-source algorithms to predict in-hospital mortality in Ugandan children with sepsis. The challenge is to first develop an algorithm using a synthetic training dataset, which will then be scored according to standard diagnostic testing criteria, and then be evaluated against a nonsynthetic test dataset. The datasets originate from admissions to six hospitals in Uganda (2017–2020) and include 3837 children, 6 to 60 months old, who were confirmed or suspected to have a diagnosis of sepsis. The synthetic dataset was created from a random subset of the original data. The test validation dataset closely resembles the synthetic dataset. The challenge should generate an optimal model for predicting in-hospital mortality. Following external validation, this model could be used to improve the outcomes for children with proven or suspected sepsis in low- and middle-income settings.

KEYWORDS: algorithms; competition; early detection and treatment; generalizability; in-hospital mortality; sepsis

Sepsis is the leading cause of death in children in low- and middle-income countries (LMICs). In 2017, there were an estimated 48.9 million cases of sepsis and 11 million sepsis-related deaths worldwide with 85% of cases and deaths occurring in LMICs (1). Children accounted for about half of these cases and nearly 3 million deaths occurred in children under 5 years old. In 2017–2020, a prospective epidemiological study was carried out in six Ugandan hospitals, examining children under 5 years old who had been admitted with a suspected or confirmed diagnosis of sepsis (2). Of note, the study found that in 3837 children 6–60 months old, 164 (4.3%) died in-hospital.

We believe that early risk stratification of children with suspected or confirmed sepsis at the time of admission may improve outcomes. Such risk stratification may also serve as a surrogate for late presentation that could inform future policies and community education initiatives for early recognition of sepsis by patients and caregivers. The "2024 Pediatric Sepsis Data Challenge" was, therefore, developed to provide an opportunity to address this gap. We propose this challenge to the pediatric critical care research community at large, and it should be of interest to data science researchers involved with risk-prediction modeling based in North American (3–5) and LMIC (6) settings.

CHALLENGE DATA SOURCE

The data for the "2024 Pediatric Sepsis Data Challenge" comes from a de-identified, curated research dataset of a study called "Smart discharges to improve post-discharge health outcomes in children: A prospective before-after

Copyright © 2024 by the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies

DOI: 10.1097/PCC.0000000000003556

Copyright © 2024 by the Society of Critical Care Medicine and Wolters Kluwer Health, Inc. All Rights Reserved.

study with staggered implementation." The original study was approved by the Mbarara University of Science and Technology (MUST) Research Ethics Committee (REC 15/10-16, approved November 28, 2017), the Uganda National Institute of Science and Technology (HS 2207, approved April 12, 2017), and the University of British Columbia/Children & Women's Health Centre of British Columbia (UBC/C&W) Research Ethics Board (REB, H16-02679, approved May 9, 2017). Caregivers/participants provided informed consent for depositing the curated dataset in an open data repository, and all research procedures were followed in accordance with the ethical standards of the MUST REC and UBC/C&W REB and the Helsinki Declaration of 1975.

The "2024 Pediatric Sepsis Data Challenge" dataset is a subgroup of 2017–2020 Ugandan epidemiological data published in 2023 (2). These prospective, multisite, observational cohort data were focused on children up to 60 months old who were admitted to any of six Ugandan hospitals with suspected or confirmed sepsis. These hospitals have a catchment population of 1.4 million children younger than 60 months old. The authors of the 2023 report have previously shown that "approximately 90% of children admitted to hospital with a proven or suspected infection in Uganda meet the International Pediatric Sepsis Consensus Conference definition for sepsis" (7); that is, "the presence of systemic inflammatory response syndrome combined with a confirmed or suspected infection" (8).

CHALLENGE DATA GENERATION

Out of the 6545 children younger than 60 months old in the 2017–2020 dataset, we selected the 3837 children 6–60 months old with 164 (4.3%) in-hospital deaths (2).

Synthetic Training Dataset

We created a synthetically generated training dataset to reduce the risk of reidentification. Even with deidentified data, there is a persistent risk of reidentification, especially in datasets with clinical variables. This problem is particularly important in the data in the proposed data challenge, since potential participants may be able to reidentify potential patients, or even the hospital itself, which are located primarily in low-resource environments with distinct characteristics. Therefore, we have taken the precaution to ensure that our data contributing sites are minimally exposed to reidentification by adding synthetic data elements.

The synthetic training set was generated from a random 70% subset of the original data with 2686 of the 3837 records. This dataset was created in R Statistical Software (R Foundation for Statistical Computing, Vienna, Austria) using the "synthpop" package (9). We used the nonparametric classification and regression tree method for synthesizing all variables. Variables were synthesized sequentially, with the first variable (the outcome variable, in-hospital mortality) synthesized via sampling with replacement, and subsequent variables synthesized conditionally on all previous variables. The number of synthetic samples generated was equivalent to the size of the training set of the original data. The rationale behind this decision was to provide data challenge participants with an environment closely resembling our actual conditions, albeit without granting direct access to the real data. This approach aims to ensure that participants build their models under conditions that closely mimic the challenges posed by our real-world data. Missing data (31% of all data cells) were also synthesized as part of this process, and rules were specified to account for missing data due to branching logic. All direct identifiers were removed to reduce the risk of reidentification, and data collected during discharge or post-discharge from the facility were not included as they cannot be used to predict in-hospital mortality. The full training dataset contains 148 variables, including clinical, social, and laboratory values (10).

Model Validation Dataset

The remaining 30% (i.e., 1151 records) withheld from original dataset will be used as a model validation dataset. Univariable distributions between the synthetic training data and test validation dataset are similar (10). The bivariate distributions between all predictor variables compared against the outcome, in-hospital mortality, were also similar, with some exceptions where a categorical variable was poorly represented.

We have evaluated two measures of distribution divergence. Taken together, these divergence statistics suggest that the synthetic training dataset and test

Copyright © 2024 by the Society of Critical Care Medicine and Wolters Kluwer Health, Inc. All Rights Reserved.

validation dataset are similar. The maximum mean discrepancy for continuous variables—in which smaller values indicate more similar datasets—between the synthetic training dataset and test validation dataset was 0.030. The Kullback-Leibler divergence (i.e., normalized value between 0 and 1 where higher value indicates more similar datasets) for continuous and categorical variables, between the synthetic training dataset and the test validation dataset was 0.915 and 0.987, respectively (10).

THE 2024 PEDIATRIC SEPSIS DATA CHALLENGE

The 2024 Pediatric Sepsis Data Challenge aims to support global participants in building skills in model development for clinical risk prediction. It is anticipated to launch on November 4, 2024, and welcomes participants from all disciplines and expertise levels, from beginners in data science to veterans of the field. Participants are asked to design a working, open-source algorithm to predict in-hospital mortality using only the provided synthetically generated dataset (10). Ideally, the model should be capable of running on a mobile device, considering environments with unreliable electrical supply and internet connectivity. The final model may eventually be used to increase the level of care for the most vulnerable children.

The challenge consists of two phases: an "unofficial" phase and an "official" phase. The first phase of the challenge serves as a testing ground for the data, scores, and submission system before the official phase begins. Participants will use the synthetically generated training dataset to train their models. This unofficial phase allows teams to start developing preliminary algorithms and the subsequent submission should include both the training code and the corresponding trained model. During this unofficial phase of the challenge, teams can submit up to five algorithms. In the official phase of the challenge, trained models will undergo evaluation on the test validation dataset, which will be kept confidential and not shared with participants at any stage. During the official phase of the challenge, teams can submit up to ten algorithms.

The challenge organizers will execute the submitted code within a contained environment on an Amazon Web Services platform. Each team will be provided with a baseline model implemented in Python. This baseline model is a random forest classifier that inputs all available features. Categorical variables are transformed into a set of binary (0/1) variables, representing the distinct values within each categorical variable.

Model predictions will be evaluated using a specifically defined metric for this challenge. The models make mortality predictions for each patient record, and the challenge score is determined based on the true positive (TP) rate (TPR) for predicting mortality given a false positive (FP) rate (FPR) of less than or equal to 0.20. We define the numbers of TPs, FPs, false negatives, and true negatives, as seen in **Table 1**. Then, regarding the scoring metrics, we consider θ as the highest decision threshold, which is the TPR when FPR is fixed at 0.20 (**Table 2**). A perfect score is 1. The

TABLE 1.
Definitions of True Positives, False Positives, False Negatives, and True Negatives

		Observed	
Outcome		**Death**	**Survivor**
Predicted	Death	True positive	False positive
	Survivor	False negative	True negative

TABLE 2.
Scoring Metrics

Scoring Metric	Definition
$FPR\theta = FP\theta/(FP\theta + TN\theta)$	(the fraction of incorrect predictions of survivors at the decision threshold of θ)
$TPR_{FPR} = TP_{FPR}/(TP_{FPR} + FN_{FPR})$	(the fraction of correct predictions of in-hospital deaths at a fixed FPR)
Score 1 = $TPR_{FRP \leq 0.20}$	(the TPR at an FPR $\leq$ 0.20)

FN = false negative, FP = false positive, FPR = false positive rate, TN = true negative, TP = true positive, TPR = true positive rate.

Copyright © 2024 by the Society of Critical Care Medicine and Wolters Kluwer Health, Inc. All Rights Reserved.

model with the highest challenge score on the full test validation dataset will be declared the winner of the 2024 Pediatric Sepsis Data Challenge.

For those interested in participating, please visit the challenge website for more details, including registration (http://www.bcchildrens.ca/globalhealth/projects-priorities/project-highlights/2024-pediatric-sepsis-data-challenge).

1 *Institute for Global Health, BC Children's Hospital and BC Women's Hospital + Health Centre, Vancouver, BC, Canada.*

2 *Department of Computer Science and Informatics, Emory University, Atlanta, GA.*

3 *Department of Anesthesia, Pharmacology & Therapeutics, University of British Columbia, Vancouver, BC, Canada.*

4 *BC Children's Hospital Research Institute, BC Children's Hospital, Vancouver, BC, Canada.*

5 *Department of Pediatrics, University of British Columbia, Vancouver, BC, Canada.*

6 *Department of Paediatrics and Child Health, Mbarara University of Science and Technology, Mbarara, Uganda.*

7 *Holy Innocents Children's Hospital, Mbarara, Uganda.*

8 *Walimu, Kampala, Uganda.*

9 *Department of Community Health, Mbarara University of Science and Technology, Mbarara, Uganda.*

10 *Maternal Newborn and Child Health Institute, Mbarara University of Science and Technology, Mbarara, Uganda.*

11 *Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA.*

12 *Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA.*

Members of the Pediatric Sepsis Data CoLaboratory (Pediatric Sepsis CoLab) can be viewed on https://wfpiccs.org/pediatric-sepsis-colab/.

Supported, in part, by grant from Grand Challenges Canada, Thrasher Research Fund, BC Children's Hospital Foundation, and Mining4Life.

Ms. Huxford's institution received funding from Grand Challenges Canada, Thrasher Research Fund, BC Children's Hospital Foundation, and Mining4Life. Dr. Businge received funding from the Pediatric Sepsis Data CoLaboratory, the World Federation of Pediatric Intensive and Critical Care Societies, the University of British Columbia, and the BC Children's Hospital Foundation. Dr. Komugisha received support for article research from Grand Challenges Canada. Dr. Tayebwa received funding from the Mbarara University of Science and Technology. Dr. Kamaleswaran received support for article research from the National Institutes of Health. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: charly.huxford@bcchr.ca

REFERENCES

1. Rudd KE, Johnson SC, Agesa KM, et al: Global, regional, and national sepsis incidence and mortality, 1990-2017: Analysis for the Global Burden of Disease Study. *Lancet* 2020; 395:200–211
2. Wiens MO, Bone JN, Kumbakumba E, et al: Mortality after hospital discharge among children younger than 5 years admitted with suspected sepsis in Uganda: A prospective, multisite, observational cohort study. *Lancet Child Adolesc Health* 2023; 7:555–566
3. Trujillo Rivera EA, Chamberlin JM, Patel AK, et al: Dynamic mortality risk predictions for children in ICUs: Development and validation of machine learning models. *Pediatr Crit Care Med* 2022; 23:344–352
4. Steif J, Brant R, Sreepada RS, et al: Prediction model performance with different imputation strategies: A simulation study using a North American ICU registry. *Pediatr Crit Care Med* 2022; 23:e29–e44
5. Heneghan JA, Walker SB, Fawcett A, et al; Pediatric Data Science and Analytics (PEDAL) subgroup of the Pediatric Acute Lung Injury and Sepsis Investigators (PALISI) Network: The Pediatric Data Science and Analytics Subgroup of the Pediatric Acute Lung Injury and Sepsis Investigators Network: Use of supervised machine learning applications in pediatric critical care medicine research. *Pediatr Crit Care Med* 2023; 25:364–374
6. Chandna A, Keang S, Vorlark M, et al: A prognostic model for critically ill children in locations with emerging critical care capacity. *Pediatr Crit Care Med* 2024; 25:189–200
7. Wiens MO, Larson CP, Kumbakumba E, et al: Application of sepsis definitions to pediatric patients admitted with suspected infections in Uganda. *Pediatr Crit Care Med* 2016; 17:400–405
8. Goldstein B, Giroir B, Randolph A; International Consensus Conference on Pediatric Sepsis: International pediatric sepsis consensus conference: Definitions for sepsis and organ dysfunction in pediatrics. *Pediatr Crit Care Med* 2005; 6:2–8
9. Nowok B, Raab GM, Dibben C: Synthpop: Bespoke creation of synthetic data in R. *J Stat Software* 2016; 74:1–26
10. Nguyen V, Huxford C, Rafiei A, et al: Data Challenges: 2023 Pediatric Sepsis Challenge. 2023. Available at: doi.org/10.5683/SP3/TFAV36. Accessed April 24, 2024

Copyright © 2024 by the Society of Critical Care Medicine and Wolters Kluwer Health, Inc. All Rights Reserved.
View publication stats