A case study on deep learning label leakage identified during a silent trial

Linda Tang[1], Michael Gao[1], Will Ratliff[1], Shems Saleh[1], Suresh Balu[1], Marshall Nichols[1], Mike Revoir[1], Emily Sterrett[2], Mark Sendak[1]
[1]Duke Institute for Health Innovation, [2]Department of Pediatrics, Duke University Hospital

Introduction:

Prospective validation of machine learning (ML) models on real-time data during a 'silent trial' is crucial to assess real-world performance and clinical relevance.[1-3] However, most published ML models are not validated prospectively.[4] For pediatric sepsis, no previously published model has been prospectively validated during a silent trial.[5-13] This may be attributed to the lack of suitable technical infrastructure, costs associated with running silent trials or support for health system leadership.[4, 14-16] In this case study, we describe the results of a silent trial evaluation of a previously developed pediatric sepsis ML model and share the steps we took to identify the underlying causes of performance discrepancies between the retrospective evaluation and prospective evaluation.

Methods:

Our team previously developed a Long short-term memory (LSTM) model to predict pediatric sepsis within the next 6 hours using commonly available EHR data elements. The sepsis outcome label was defined according to the Weiss definition.[17] The model was trained on 17,491 encounters between 11/1/2016 and 12/31/2020 at Duke University Health System (DUHS), and 464 (2.65%) encounters met the sepsis definition. The temporal validation cohort consists of 6,545 encounters from 1/1/2021 to 6/30/2022. Despite class imbalance, the LSTM model achieved robust predictive performance on the retrospective test set and on the temporal validation cohort (AUROC of 0.936 and 0.937, AUPRC of 0.440 and 0.405, respectively). The model was integrated into the operational EHR system at DUHS. A custom-built database extracts the most recent data from EPIC every 15 minutes, which provides near real-time data for patients who are currently hospitalized. The model runs every 15 minutes on all current encounters at DUHS and generates a notification if a patient's risk score exceeds the pre-set threshold of 0.85. During the 2-month silent trial, the model outputs were not sent to bedside clinicians. Instead, the notifications were sent to an internal HIPAA-compliant message channel where the model development team could track the alarm volume and resolve technical issues. Additional analyses were designed to investigate the observed performance discrepancies.

Results:

The model ran on more than 1,475 encounters during the silent trial. The model generated approximately 30 alarms per day, which was much higher than the expected 2 alarms per day based on retrospective performance. In addition, our team noticed that the model fired an alarm for almost all patients in the emergency department within their first hour of arrival. During our investigation of performance discrepancies, we observed that in the retrospective modeling, we truncated the data for septic encounters at the time of meeting the real-time Weiss phenotype. However, we did not truncate any encounter's data when running the model in real-time. As a result, the encounter length for septic patients is much longer during the silent trial than in retrospective modeling. The average length of stay of non-septic patients is 77.3 hours, and the average length of stay of septic patients is 459.6 hours before truncation, but 39.6 hours after truncation. While training the model, we set the maximum length for all encounters to be 168 hours. If an encounter was less than 168 hours long, zeros were used to pad the sequence to achieve a length of 168. However, as we normalized the outputs from the LSTM using layer normalization, the shorter sequences (i.e. with more zeros) got a smaller mean. Due to this difference in means, the model learned to associate shorter encounters with sepsis and longer encounters with non-sepsis. After this realization, we retrained the LSTM model without layer normalization using the same hyperparameters, which resulted in an AUROC of 0.782 and AUPRC of 0.01 on the retrospective cohort. This suggests that the model's predictive power was inflated due to label leakage introduced through layer normalization.

Discussion:

Prospective validation helped our team detect an instance of label leakage despite the model's robust performance on the retrospective test set and temporal validation set. Completing temporal validation was not sufficient to help our team identify the label leakage because all data (formatted into an hourly model matrix) were fed into the model at once, whereas prospective validation requires data to be fed into the model at pre-specified prediction cadences (e.g. new data added every hour), which more closely mimics the model's clinical use case. Identifying this issue before integrating the tool into clinical workflow helped our team avoid causing alarm fatigue and damaging end-user trust. It is possible that other published deep-learning models may also suffer from similar subtle instances of label leakage. We strongly recommend other researchers to prospectively validate their ML models prior to utilization in patient care.

References:

1. Kwong JCC, Erdman L, Khondker A, et al. The silent trial - the bridge between bench-to-bedside clinical AI applications. *Front Digit Health.* 2022;4:929508. Published 2022 Aug 16. doi:10.3389/fdgth.2022.929508
2. McCradden MD, Anderson JA, A Stephenson E, et al. A Research Ethics Framework for the Clinical Translation of Healthcare Machine Learning. *Am J Bioeth.* 2022;22(5):8-22. doi:10.1080/15265161.2021.2013977
3. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care [published correction appears in Nat Med. 2019 Oct;25(10):1627]. *Nat Med.* 2019;25(9):1337-1340. doi:10.1038/s41591-019-0548-6
4. Sendak MP, D'Arcy J, Kashyap S, et al. A path for translation of machine learning products into healthcare delivery. *European Medical Journal.* Published February 4, 2020.
5. Cabrera-Quiros L, Kommers D, Wolvers MK, et al. Prediction of Late-Onset Sepsis in Preterm Infants Using Monitoring Signals and Machine Learning. *Crit Care Explor.* 2021;3(1):e0302. Published 2021 Jan 27. doi:10.1097/CCE.0000000000000302
6. Helguera-Repetto AC, Soto-Ramírez MD, Villavicencio-Carrisoza O, et al. Neonatal Sepsis Diagnosis Decision-Making Based on Artificial Neural Networks. *Front Pediatr.* 2020;8:525. Published 2020 Sep 11. doi:10.3389/fped.2020.00525
7. Huang B, Wang R, Masino AJ, Obstfeld AE. Aiding clinical assessment of neonatal sepsis using hematological analyzer data with machine learning techniques. *Int J Lab Hematol*. 2021;43(6):1341-1356. doi:10.1111/ijlh.13549
8. Kamaleswaran R, Akbilgic O, Hallman MA, West AN, Davis RL, Shah SH. Applying Artificial Intelligence to Identify Physiomarkers Predicting Severe Sepsis in the PICU. *Pediatr Crit Care Med.* 2018;19(10):e495-e503. doi:10.1097/PCC.0000000000001666
9. Lamping F, Jack T, Rübsamen N, et al. Development and validation of a diagnostic model for early differentiation of sepsis and non-infectious SIRS in critically ill children - a data-driven approach using machine-learning algorithms. *BMC Pediatr.* 2018;18(1):112. Published 2018 Mar 15. doi:10.1186/s12887-018-1082-2
10. Masino AJ, Harris MC, Forsyth D, et al. Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PLoS One.* 2019;14(2):e0212665. Published 2019 Feb 22. doi:10.1371/journal.pone.0212665
11. Ehwerhemuepha L, Heyming T, Marano R, et al. Development and validation of an early warning tool for sepsis and decompensation in children during emergency department triage. *Sci Rep.* 2021;11(1):8578. Published 2021 Apr 21. doi:10.1038/s41598-021-87595-z
12. Le S, Hoffman J, Barton C, et al. Pediatric Severe Sepsis Prediction Using Machine Learning. *Front Pediatr.* 2019;7:413. Published 2019 Oct 11. doi:10.3389/fped.2019.00413
13. Scott HF, Colborn KL, Sevick CJ, et al. Development and Validation of a Predictive Model of the Risk of Pediatric Septic Shock Using Data Known at the Time of Hospital Arrival. *J Pediatr.* 2020;217:145-151.e6. doi:10.1016/j.jpeds.2019.09.079
14. Sendak MP, Balu S, Schulman KA. Barriers to Achieving Economies of Scale in Analysis of EHR Data. A Cautionary Tale. *Appl Clin Inform*. 2017;8(3):826-831. Published 2017 Aug 9. doi:10.4338/ACI-2017-03-CR-0046
15. Sendak M, Gao M, Nichols M, Lin A, Balu S. Machine Learning in Health Care: A Critical Appraisal of Challenges and Opportunities. *EGEMS (Wash DC)*. 2019;7(1):1. Published 2019 Jan 24. doi:10.5334/egems.287
16. Sendak M, Gulamali F, Balu S. Overcoming the activation energy required to unlock the value of AI in Healthcare. *National Bureau of Economic Research.* Published September 23, 2022.
17. Weiss SL, Peters MJ, Alhazzani W, et al. Surviving Sepsis Campaign International Guidelines for the Management of Septic Shock and Sepsis-Associated Organ Dysfunction in Children. *Pediatr Crit Care Med*. 2020;21(2):e52-e106. doi:10.1097/PCC.0000000000002198