Towards Deploying Predictive Models for Maternal Health

Cedric Diggory[1]
[1] *Hogwarts School of Witchcraft and Wizardry*

Background. Maternal mortality in the US has been a growing problem; the maternal mortality rate has increased steadily over the past few years from 20.1 per hundred thousand live births in 2019 to 32.9 in 2021[1], one of the highest rates across high-socioeconomic status countries.[2] Analysis of data from the Maternal Mortality Review Committees found that 80% of maternal deaths are preventable[3]. An early warning system could identify patients at risk and improve clinical outcomes.

Prior work by our institution identified and predicted maternal sepsis but focused on only one birthing center, had a high proportion of false positives during and shortly after delivery, and did not measure the potential of or anti-black algorithmic bias. In this work, we describe the effort to operationalize this maternal early warning system, addressing the above shortcomings by including multiple birthing centers , measuring model performance by racial subgroups, and integrating the model into a real-time dashboard that will be used in clinical care for patients in maternal units.

Methods. This work uses data for 70,544 encounters from two hospitals within one health system. The model was trained on data from Jan 2016 to Dec 2021. The model was trained on 2016–2020 data and tested on 2021 data. Data quality was evaluated by a clinician and data scientist team utilizing a data quality assurance framework to evaluate completeness, conformance, and plausibility of lab and structured nurse documented values.[4] The input data included both static data at admission (age, race, gender, ethnicity, hospital and ward location, prenatal encounters, and prior comorbidities) and real-time data (labs, vitals, orders, medications, and blood loss).

Every hour, the model predicts whether the patient will meet the phenotype for sepsis within the next four hours. In this context, sepsis is defined as {temp<36C, temp>38C, or positive blood culture} and {2 of 5: SBP<100, HR>120, RR>24, O_2<92%, WBC>17} within a two hour window. We used a Logistic Regression model, which outperformed other models (including LSTM and XGBoost) in initial experiments. The evaluation metrics presented in this work focuses on alert-level precision and encounter-level recall. Such new metrics deliver more realistic indications of clinical utility than alert-level recall. We report performance of the model both overall and broken down by race.

Results. Table 1 shows the performance of the model. One immediate observation is that metrics like AUROC and AUPRC – which look at average performance across all thresholds – diverge from metrics like precision and recall for a fixed threshold. This divergence can occur because integrating over the whole curve of performance obscures how well the given tool would do at a given threshold (as is the case when in use). Although AUROC and AUPRC suggest the presence of anti-black algorithmic bias when integrating across all thresholds, we see that for given thresholds of interest, the black and white patients have similar detection rates.

Quality assurance and spot checking model performance helped identify areas where domain expertise improved the model. For instance, during chart review to validate the model, the clinician noticed many sepsis false positives were triggered by abnormal vitals in the hours after delivery. Adding features related to post-delivery recovery helped reduce false positives and improve the model.

Table 1: Overall (and per-race) performance of the maternal sepsis model.

Cohort	n_enc	n_pred	AUROC	AUPRC (enc-level)	Recall @ 20% Ovr Prec	Prec @ 20% Ovr Recall
Total	12,083	545,438	0.94	0.12	0.206	0.203
White	5,304	249,732	0.949	0.132	0.178	0.175
Black	4,397	188,938	0.905	0.082	0.184	0.189
Asian	684	35,078	0.953	0.162	0.234	0.308

Conclusion. Our close collaboration between clinicians and data scientists on data quality, feature selection, and model evaluation generated significant insights that influenced the development and performance of the Maternal Early Warning System (MEWS). Next steps include clinical validation of the model and integration into the workflow of clinicians on the maternal floors of the hospital.

---

[1] Centers for Disease Control and Prevention. (2023, March 16). *Maternal mortality rates in the United States, 2021*. Centers for Disease Control and Prevention. Retrieved April 7, 2023, from https://www.cdc.gov/nchs/data/hestat/maternal-mortality/2021/maternal-mortality-rates-2021.htm#Table

[2] Commonwealth Fund. (2022, December 1). *The U.S. maternal mortality crisis continues to worsen: An international comparison*. U.S. Maternal Mortality Crisis Continues to Worsen. Retrieved April 7, 2023, from https://www.commonwealthfund.org/blog/2022/us-maternal-mortality-crisis-continues-worsen-international-comparison

[3] Centers for Disease Control and Prevention. (2022, September 19). *Four in 5 pregnancy-related deaths in the U.S. are preventable*. Centers for Disease Control and Prevention. Retrieved April 7, 2023, from https://www.cdc.gov/media/releases/2022/p0919-pregnancy-related-deaths.html

[4] Sendak, M., Sirdeshmukh, G., Ochoa, T., Premo, H., Tang, L., Niederhoffer, K., Reed, S., Deshpande, K., Sterrett, E., Bauer, M., Snyder, L., Shariff, A., Whellan, D., Riggio, J., Gaieski, D., Corey, K., Richards, M., Gao, M., Nichols, M., … Balu, S. (2022, December 31). *Development and validation of ML-DQA – a machine learning data quality assurance framework for Healthcare*. Proceedings of the 7th Machine Learning for Healthcare Conference. Retrieved April 7, 2023, from https://proceedings.mlr.press/v182/sendak22a.html