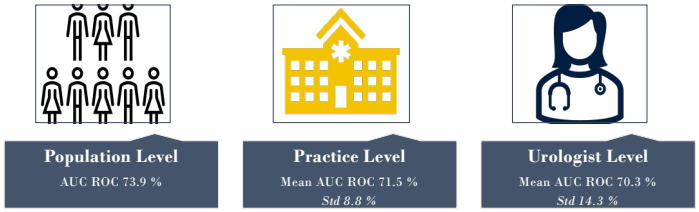
Physicians Using Machine Learning Predictions Observe Worse Performance than Reported

Erkin Ötleş[1, a], Johannes Allgaier[2, b], Karandeep Singh[1, c]
[1] University of Michigan, U.S.A, [2] University of Würzburg, Germany
[a] eotles@med.umich.edu, [b] johannes.allgaier@uni-wuerzburg.de, [c] kdpsingh@med.umich.edu



Background

Accompanying advances in machine learning (ML), there has been a proliferation of predictive models being developed for and implemented in clinical practice. Despite the overall interest in developing and implementing models, there have been serious issues raised with the **quality of evaluations being produced**, especially for proprietary models [Wong et al., 2021, Singh et al., 2020, Lyons et al., 2023]. Widely implemented models, such as the Epic Sepsis Model and Epic Deterioration Index, are deployed across health systems, being used to generate predictions for many different types of patients and physicians.

Methods

We propose an evaluation of an implemented model on real patient data on different levels: Population (includes all treated patient records), practice (includes all patients treated in that practice), urologist (includes all patients treated by this urologist). The evaluation focuses on the Michigan Urological Surgery Improvement Collaborative (MUSIC) radical prostatectomy outcome prediction model applied to the urology practices that are a part of the MUSIC collaborative at the University of Michigan [Ötleş, 2022]. The dataset contains 2400 patient records from 227 urologists which are employed in a total of 42 practices. We only included practices and urologists with at least 10 treated patients and calculated the observed Area Under the Receiver Operator Curve (AUC ROC) Score and Brier Score Loss (BS) on each level, population, practice and urologist.

Results

Surprisingly, the mean performances at more granular levels (practice, urologist) did not match the performance at population level. 15 of 26 practices (57.7%) and 43 of 72 urologists (59.7%) observed **a worse AUC ROC** than indicated at population level. Also at population level, the AUC ROC is 73.9 %, at practice level the mean is 71.5 % (*std 8.8 %*) and at urologist level the mean is only 70.3 % (*std 14.3 %*). Within a practice with at least 2 urologists, 61 % observe a worse AUC ROC on their urologist level than the practice on its practice level.

For **BS**, the performance gaps between observer levels are smaller. At population level, BS is 20.2 %, at practice level mean BS is 20.8 % (*std 3.2 %*) and at urologist level, mean BS is 20.7 % (*std 4.4 %*). 14 out of 26 practices (53.8 %) and 41 out of 72 (56.9 %) urologists observe a worse BS than at population level.
The more granular the level, the larger the standard deviation and the worse performance metrics (AUC ROC and BS) are observed.

Conclusion

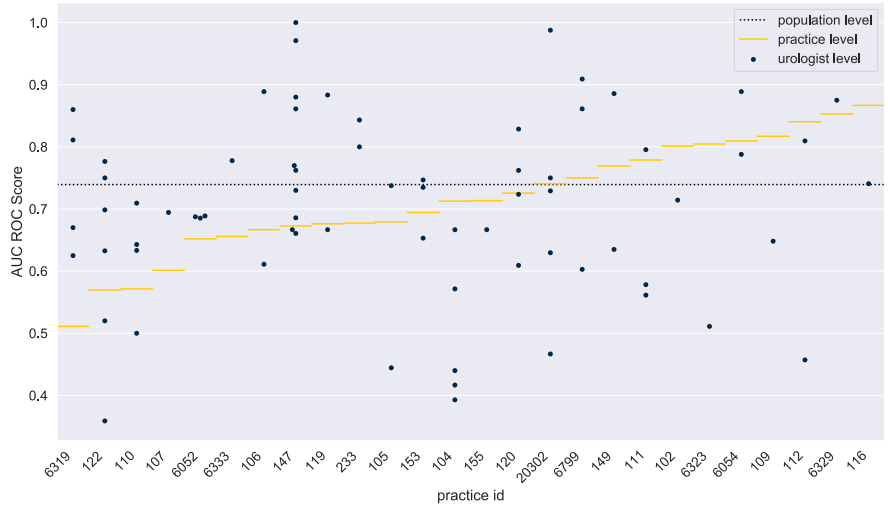
In this study we show that the performance of a ML system perceived on observer level (urologist, practice) is less than the performance on the entire population. We think there might be hidden features in the dataset, i. e., the observers (practices, urologists) themselves. It is also possible that certain patient groups (e.g., severe cases) are only seen by a particular physician, but the feature space of these patients is not covered granularly enough by the model. It is crucial to understand the performance of these models as seen by individual observers to improve the performance of ML systems used in healthcare. The population AUROC / BS is never observed by observers who only see a subset of patients who generally are more homogenous than the general population. This may reduce the observed discriminative abilities of the model.

References

- Wong, A., Ötleş, E., Donnelly, J.P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C. and Ghous, M., 2021. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, *181*(8), pp.1065-1070.
- Singh, K., Valley, T.S., Tang, S., Li, B.Y., Kamran, F., Sjoding, M.W., Wiens, J., Otles, E., Donnelly, J.P., Wei, M.Y. and McBride, J.P., 2021. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. Annals of the American Thoracic Society, 18(7), pp.1129-1137.
- Lyons, P.G., Hofford, M.R., Sean, C.Y., Michelson, A.P., Payne, P.R., Hough, C.L. and Singh, K., 2023. Factors Associated With Variability in the Performance of a Proprietary Sepsis Prediction Model Across 9 Networked Hospitals in the US. JAMA Internal Medicine.
- Ötleş, E., Denton, B.T., Qu, B., Murali, A., Merdan, S., Auffenberg, G.B., Hiller, S.C., Lane, B.R., George, A.K. and Singh, K., 2022. Development and validation of models to predict pathological outcomes of radical prostatectomy in regional and national cohorts. The Journal of Urology, 207(2), pp.358-366.

Observed Performances on Population-, Practice-, and Urologist Level

(A) AUC ROC Score - Higher means better



(B) Brier Score Loss - Lower means better

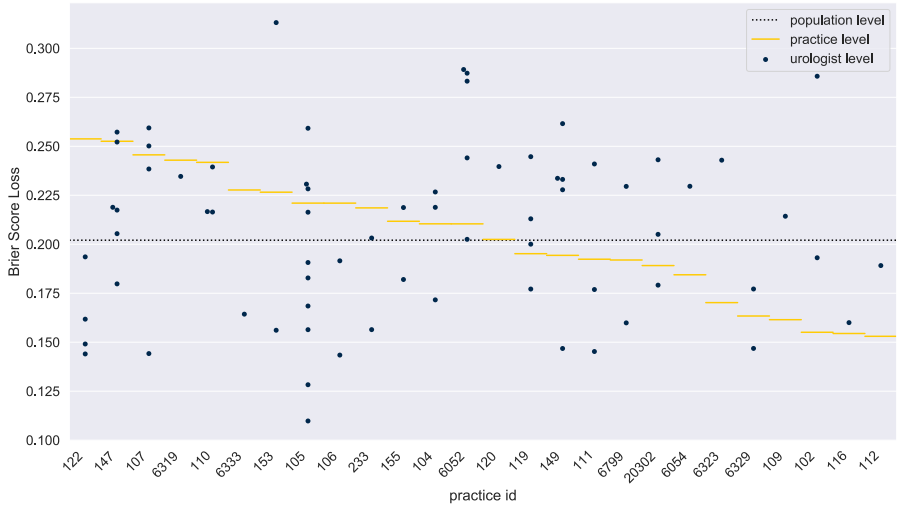


Figure.
Vertically stacked green circles indicate the observed performance of urologists with at least 10 patient treatments in this practice. Red horizontal bars indicate the observed performance at practice level calculated using all patients treated in this practice with a minimum of 10 patients per practice.