Towards standardized quality control for AI systems in critical care

In intensive care units (ICUs), physiological parameters of patients are constantly monitored and combined with laboratory data and electronic health records to facilitate optimal and timely decision making, e.g., in cases of sudden deterioration. The wealth and inhomogeneity of the available data, however, makes it difficult for healthcare professionals to take all facets of a patient's condition into account. Artificial intelligence (AI) and machine learning (ML) based systems hold great promise to support clinical decision making in this context. Trained on historical data with known outcome, such systems can predict a patient's future clinical trajectory from the multivariate and multimodal data stream. Recently, retrospective studies have demonstrated the capability of ML models to predict the emergence of acute kidney injury (Tomašev et al., 2019), a variety of complications in cardiothoracic surgery (Meyer et al., 2018), as well as of lethal courses in Covid-19 pneumonia (Lichtner et al., 2020) with high accuracy. The prospective adoption of ML models in clinical practice, however, raises additional questions regarding the fairness, robustness, certainty, and comprehensibility of their decisions on a single case basis. Models should moreover be robust to moderate changes of the input data due to measurement errors, missing data, outliers as well as distributional shifts that can occur when transferring learned models to new clinical sites. It is also desirable that ML models provide well-calibrated estimates of the uncertainty of their individual predictions, as such information can be used to implement different (liberal vs. conservative) treatment strategies. A systematic protocol to benchmark the quality of ML models in critical care along these dimensions has not been provided yet.

The proposed project will carry out first steps towards a standardized quality assessment of machine learning approaches in critical care. To this end, the project will define a suite of benchmarks including quality criteria, reference problems, reference data, and reference implementations of ML models. The successful applicant will be supervised by Dr. Stefan Haufe (new professorship on machine learning and inverse problems at TU Berlin) and closely collaborate with the Institute of Medical Informatics at Charité. At a later stage, a larger community around the topic should be build (e.g. through public data analysis challenges). The project will primarily use large public data sets such as MIMIC III/IV (Johnson et al., 2016), the AmsterdamUMCdb, and HiRID (Hyland et al., 2020). Established and novel clinically relevant learning problems will be defined, such as the prediction of mortality, the prediction post-operative delir, and the prediction of individual treatment efficacies. Data from the different centers will be cleaned and harmonized. Reference prediction models from the literature will be implemented and their generalizability will be studied. Perturbed versions of the data will be created to study the robustness of the predictions to outliers, missing data, noise, and distributional shifts.

Approaches to obtain uncertainty estimates for predictions will be implemented and validated. Recommendations on how to achieve balanced performance according to different criteria will be formulated. Eventually, novel prediction models based on state-of-the-art neural network architectures (e.g., recursive, invertible, self-calibrated, attention-based) should be developed and validated.

Principal Investigator

- Stefan Haufe (PTB 8.44, TU Berlin and Charité)

PhD student

- Rustam Zhumbagetov (University supervisor: Prof. Stefan Haufe, TU Berlin)

References

Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019 Aug;572(7767):116-119.

Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, Stamm C, Hofmann T, Falk V, Eickhoff C. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med. 2018 Dec;6(12):905-914.

Lichtner G, Balzer F, Haufe S, Giesa N, Schiefenhövel F, Schmieding M, Jurth C, Kopp W, Akalin A, Schaller SJ, Weber-Carstens S, Spies C, von Dincklage F. Predicting lethal courses in critically ill COVID-19 patients using a machine learning model trained on patients with non-COVID-19 viral pneumonia. 2020. Submitted.

Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035.

Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med. 2020 Mar;26(3):364-373.