BEHRT-HF: an interpretable transformer-based, deep learning model for prediction of incident heart failure

S. Rao, Y. Li, R. Ramakrishnan, A. Hassaine, D. Canoy, Y. Zhu, G. Salimi-Khorshidi, K. Rahimi

University of Oxford, Oxford, United Kingdom

Funding Acknowledgement: Type of funding source: Private grant(s) and/or Sponsorship. Main funding source(s): National Institute for Health Research, Oxford Martin School, Oxford Biomedical Research Centre

Background/Introduction: Predicting incident heart failure has been challenging. Deep learning models when applied to rich electronic health records (EHR) offer some theoretical advantages. However, empirical evidence for their superior performance is limited and they remain commonly uninterpretable, hampering their wider use in medical practice.

Purpose: We developed a deep learning framework for more accurate and yet interpretable prediction of incident heart failure.

Methods: We used longitudinally linked EHR from practices across England, involving 100,071 patients, 13% of whom had been diagnosed with incident heart failure during follow-up. We investigated the predictive performance of a novel transformer deep learning model, "Transformer for Heart Failure" (BEHRT-HF), and validated it using both an external held-out dataset and an internal five-fold cross-validation mechanism using area under receiver operating characteristic (AUROC) and area under the precision recall curve (AUPRC). Predictor groups included all outpatient and inpatient diagnoses within their temporal context, medications, age, and calendar year for each encounter. By treating diagnoses as anchors, we alternatively removed different modalities (ablation study) to understand the importance of individual modalities to the performance of incident heart failure prediction. Using perturbation-based techniques, we investigated the importance of associations between selected predictors and heart failure to improve model interpretability.

Results: BEHRT-HF achieved high accuracy with AUROC 0.932 and AUPRC 0.695 for external validation, and AUROC 0.933 (95% CI: 0.928, 0.938) and AUPRC 0.700 (95% CI: 0.682, 0.718) for internal validation. Compared to the state-of-the-art recurrent deep learning model, RETAIN-EX, BEHRT-HF outperformed it by 0.079 and 0.030 in terms of AUPRC and AUROC. Ablation study showed that medications were strong predictors, and calendar year was more important than age. Utilising perturbation, we identified and ranked the intensity of associations between diagnoses and heart failure. For instance, the method showed that established risk factors including myocardial infarction, atrial fibrillation and flutter, and hypertension all strongly associated with the heart failure prediction. Additionally, when population was stratified into different age groups, incident occurrence of a given disease had generally a higher contribution to heart failure prediction in younger ages than when diagnosed later in life.

Conclusions: Our state-of-the-art deep learning framework outperforms the predictive performance of existing models whilst enabling a data-driven way of exploring the relative contribution of a range of risk factors in the context of other temporal information.