Abstract:

The abundance, diversity, and complexity of electronic healthcare records (EHR) allow the building of principled prognostic models to support informed decision making by professionals and individuals. Given the amount of data, building a data-driven machine learning (ML) approach is a natural choice. Although this has been a popular approach in recent years, explainability and operationalizability remain two critical issues. Explainability requires the decision made by the ML model to be accessible to practitioners, end-users and stakeholders, while operationalizability requires the model to be applicable to real-life real-time decision making under imperfect information. This project aims to investigate existing models, both statistical and ML for explainability and operationalizability and develop tools that are explainable and operationalizable alongside being accurate, calibrated, stable, adaptable, transferable and able to quantify uncertainty.

Introduction:

Populations around the world are ageing rapidly. It is predicted that by 2050 the proportion of people who are aged 60+ years will nearly double from 12% (2015 estimate) to 22% [3]. Ageing populations drive increasing frailty, multimorbidity and polypharmacy, all responsible for higher mortality, diminished quality of life and increased health and social care needs [4]. Better prediction of individual prognosis (e.g., of mortality, hospital admission and institutionalization), allows targeted early intervention potentially leading to better quality of later life and economic benefit. The availability of electronic health care data facilitates this by providing unprecedented detail of the longitudinal health trajectory of an individual, but more sophisticated methods are needed to deal with the incompleteness of the data, to extract causal reasoning beyond simple association, and to adapt to the dynamic nature of individual and population health.

Research Challenge:

Due to the abundance of routine healthcare data, data-driven ML modelling has gained popularity in deriving new insights that might be missed by traditional modelling. However, there are open challenges, and both these approaches have elements that should be brought together to develop risk prediction tools that are more explainable and operationalizable. For example, from a ML perspective, the literature often focusses on predicting a binary outcome rather than the time-to-event models favoured in biostatistics which better account for routine healthcare data usually being censored. Censoring is often also informative (e.g., an individual may be lost to follow up due to death, and competing mortality risk influences estimated risk - you cannot get a lung transplant if you have already died of lung cancer). Furthermore, the intricate nature of multiple risks interacting with each other, e.g., through repeated outcomes etc. are often ignored for simplicity. On the other hand, a similar situation arises when a model is trusted throughout time (i.e., periods) when it is well understood that the characteristics of the model might have changed due to distributional shifts inherent to the nature of the data (e.g., risk of death from COVID-19 before and after vaccine is very different) and thus a dynamic model, that is designed for streaming data that is adaptable is more sensible than a conventional biostatistical static model that has been trained once. Recent development in causal machine learning [2] has added another dimension to risk prediction by capturing causation beyond association particularly in the presence of nonlinearity and high dimensionality. This contributes to better explainability through counterfactual reasoning, and better operationalizability through interventional paradigms. Therefore, we address the research challenge of assessing and designing explainable and operationalization risk prediction tools for later life risk prediction by combining strengths from established statistical models and more contemporary ML model.

Data & Methodology:

The project will be using longitudinal survey data from the English Longitudinal Study of Ageing (https://www.elsa-project.ac.uk) and longitudinal routine healthcare data from the SAIL Databank (https://saildatabank.com). The objectives of the projects are to perform a review of existing methods of risk prediction tools in the context of ageing for their explainability and operationalizability, implement these methods in the context of ELSA and SAIL, and develop tools combining the strengths of these approaches. The methods include a) standard statistical models such as multi-state models (MSM) with time-varying covariates [5], b) data-driven models such as Random Survival Forests (RSF) [6] with competing risks and time-varying covariates, c) causal models such as Causal Inference over Mixture (CIM), d) dynamic models such as DeepHit [1] that considers competing risks, dynamic prediction and uncertainty quantification, etc.

RRI/Ethical Considerations:

Designing better prognosis models is of utmost importance in facilitating targeted care for an ageing population to improve quality of life. Although several models have been presented in the literature, the clinical translation of these models is often limited due to lack of transparency and therefore mistrust by clinical users. Improving these models, therefore, will benefit patients, practitioners and public in general. The project will be held in collaboration with the Advanced Care Research Centre and we will collaborate with the established Patient and Public Involvement and Engagement to enforce these elements of responsible innovation. The project will use ethical protocols set up by the data providers.

Expected Outcome & Impact:

The expected outcome of the project is an assessment of the explainability and operationalizability of existing predictive modelling in health and care particularly in the context of hospital admission, mortality and institutionalization, and developing innovative tools from a causal perspective that are robust, accurate, well calibrated, transparent, trustworthy, adaptable to the dynamic environment, and that can be deployed for clinical use in the future. The study will build a case for using state-of-the-art ML tools in risk prediction to the end-user to facilitate clinical translation.

References:

[1] https://doi.org/10.1609/aaai.v32i1.11842
[2] 10.1038/s41591-024-02902-1
[3] https://www.who.int/news-room/fact-sheets/detail/ageing-and-health
[4] https://doi.org/10.1016/S2666-7568(24)00007-2
[5] 10.1023/a:1009672031531
[6] https://doi.org/10.1186/s12874-021-01375-x