Name: Sohan Seth
e-mail: sohan.seth@ed.ac.uk
website: https://sohanseth.github.io
affiliation: CSE
School: School of Informatics

Name: Bruce Guthrie
e-mail: bruce.guthrie@ed.ac.uk
website: https://www.ed.ac.uk/profile/bruce-guthrie
affiliation: CMVM
School: Usher Institute

Name: Ewen Harrison
e-mail: ewen.harrison@ed.ac.uk
website: https://www.ed.ac.uk/profile/ewen-harrison
affiliation: CMVM
School: Usher Institute

External Partner:
Contact Name:
Contact e-mail address:
Organization name:
Organization website:
Role of external partner:

If no EP, potential ones to co-develop:

Primary theme alignment: AI for Biomedical and Health Informatics
Project title: **Operationalizable clinical risk predictions using machine learning-driven multi-state models**

Summary (400 words): Better prediction of individual prognosis (e.g., mortality, hospital admission/readmission and institutionalization), allows targeted early intervention potentially leading to a better quality of life and economic benefit. The availability of electronic healthcare records facilitates this by providing unprecedented detail of the longitudinal health trajectory of an individual. However, more sophisticated methods are needed to deal with the incompleteness of the data, the existence of multiple competing outcomes, and the dynamic relationship between covariates and outcomes. Furthermore, these models should be explainable and operationalizable to ensure their clinical translation. This project aims to develop machine learning-based multi-state models to predict multiple outcomes simultaneously and assess their accuracy, efficiency, stability, adaptability, transferability, and capacity to quantify uncertainty. The project will be using longitudinal survey data from the English Longitudinal Study of Ageing (https://www.elsa-project.ac.uk), longitudinal routine healthcare data from the SAIL Databank (https://saildatabank.com) and the critical care dataset from the ISARIC4C study (https://isaric4c.net).

Description (1000 words):

**Abstract:**

The abundance, diversity, and complexity of Electronic Health Records (EHR) allow the building of principled prognostic models to support informed decision-making by health professionals. Building a data-driven machine learning (ML) approach is a natural choice given the amount of data. Although this has been a popular approach in recent years, these models are often simplified to ignore 'competing risks' that are commonplace in EHR data. This project aims to develop ML-based multi-state models (MSM) [3] to predict multiple related outcomes in longitudinal healthcare data and to address two critical issues of explainability and operationalizability. Explainability requires the decision made by the ML model to be accessible to practitioners, end-users and stakeholders, while operationalizability requires the model to apply to real-life real-time decision-making under imperfect information.

**Introduction:**

Predicting the risk of adverse outcomes is critical in clinical informatics to develop early intervention. For example, in later-life care [6], predicting the risk of mortality, hospital admission and institutionalization, is crucial for maintaining a better quality of life, while in critical care, predicting the risk of one or more hospital or ICU readmissions is critical for designing better care pathways. In both cases, better prediction of individual prognosis allows targeted early intervention potentially leading to better quality of later life and economic benefit. The availability of EHR data facilitates this by providing unprecedented detail of the longitudinal health trajectory of an individual. Still, more sophisticated methods are needed to deal with the incompleteness of the data, the existence of multiple competing outcomes, and the possibly dynamic relationship between covariates and outcomes.

**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. These models, however, often make simplifying assumptions. For example, from a machine learning perspective, the literature often focuses on predicting a binary outcome rather than the time-to-event models favoured in biostatistics which better account for routine healthcare data usually having censoring. Censoring is often 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. A natural choice of addressing these aspects of the data is to use a MSM with time-varying covariates and effects. Multi-state models have recently been used in conjunction with machine learning to address the limitations posed by linearity and proportionality, but this area remains relatively less explored, and the existing models remain to be applied to large clinical datasets often due to the associated computational burden. This project aims to contribute to this growing literature by developing efficient ML-driven MSM [1] and apply these models EHR with a particular focus on explainability and operationalizability. We will additionally explore making these models computationally efficient and extracting causal understanding beyond association [2].

**Data & Methodology:**

The project will be using longitudinal survey data from the English Longitudinal Study of Ageing (https://www.elsa-project.ac.uk), longitudinal routine healthcare data from the SAIL Databank (https://saildatabank.com), and critical care dataset from the ISARIC4C study (https://isaric4c.net). The objectives of the projects are to perform a review of existing methods of risk prediction tools in the context of ageing and critical care with a particular focus on their explainability and operationalizability, implement these methods in the context of ELSA, SAIL and ISARIC, and to develop ML-base MSM combining the strengths data-driven approaches and the finesse of epidemiology. The project will also explore other baseline models e.g., Random Survival Forests (RSF) [4], Causal Inference over Mixture (CIM) [5] etc.

**RRI/Ethical Considerations:**

Designing better prognosis models is of utmost importance in facilitating targeted care 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 novel machine learning driven multi-state models that are applicable in practice particularly in the context of hospital admission, mortality and institutionalization, and assessing their robustness, accuracy, calibration, transparency, trustworthiness, adaptability to dynamic environment. 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/10.1007/s10462-023-10681-3
[2] 10.1038/s41591-024-02902-1
[3] 10.1023/a:1009672031531
[4] https://doi.org/10.1186/s12874-021-01375-x
[5] https://doi.org/10.1007/s10994-022-06159-y
[6] https://doi.org/10.1016/j.archger.2019.103974