Project title: Causal Discovery for Large-Scale Analysis of Patient Trajectories from Health Data

Project reference: DT4H_07_2023

1st supervisor: Dr Yu Luo - Department of Mathematics

2nd supervisor: Dr David Watson - Department of Informatics

Aim of the project

Understanding the progression of chronic diseases is important to inform early diagnosis, personalize care and ensure effective healthcare system management. Data from clinical and administrative systems have the potential to advance this understanding, but traditional methods for modelling disease progression are ill-suited to many healthcare datasets, where samples may be collected at irregular intervals and cohorts are large and heterogeneous.

The impact of treatment intervention, and the discovery of optimal adaptive treatment strategies (ATSs), is of great importance in primary and clinical care. Most optimal causal discovery approaches have been attempted in relatively small-scale settings. In this project, we will develop digital twins for ATS discovery using Bayesian and machine learning approaches, which scale better to large, complex datasets. These methods promise to have a major impact on clinical decision making, opening the door to data□driven personalized care.

Project description

Understanding the progression of chronic diseases like chronic obstructive pulmonary disease is crucial for early diagnosis, personalized care, and efficient health system management. Clinical and administrative data have the potential to enhance our knowledge in this area. However, traditional methods for modeling disease progression are not well-suited to analyzing high-dimensional data collected irregularly.

Many severe illnesses can be attributed to inadequate and inconsistent follow-up healthcare. To gain insights into the root causes of severe and preventable illnesses, we must not only assess the number of patients with severe conditions but also understand their "care paths" or "trajectories" leading up to their severe illness.

Both private and public healthcare systems maintain longitudinal digital health records, including data from electronic health records, health systems, and mobile health applications. These data can help inform clinical and public health decisions by studying individual patient trajectories. However, data is typically recorded only during patient□provider interactions, leading to irregularly spaced observations, and the patterns of these clinical interactions vary from patient to patient. Since the underlying disease progression s not directly observable, inferential methods are necessary to determine this latent progression. Analysing such data, especially in large cohorts with frequent observations, is a complex task.

The overarching goal of this research project is to advance statistical methodologies for the

discovery of optimal adaptive treatment in chronic diseases. The student involved in this project will achieve the following specific aims:

1. To develop a modelling framework that accounts for non-random observation time points. In medical research, observation times can provide valuable insights into a patient's illness. The model may include a state-dependent measurement intensity to capture this.

2. To develop and enhance digital twins for optimal treatment causal discovery, including both Bayesian methods and machine learning approaches, especially for large-scale analyses. Bayesian methods are a particularly active area of research for identifying optimal ATSs. These methods are appealing because they enable comprehensive propagation and consideration of inferential uncertainty, even in the context of large-scale data analysis.

3. To apply the extended model to a longitudinal drug efficacy study in multiple sclerosis and rheumatoid arthritis. This will allow clinicians to optimize personalized care strategies by using patient trajectories for dynamic treatment decisions.

The ideal candidate will have a background in probability and statistics, as well as some programming experience. Previous experience working with healthcare data is desirable but not required.