Project title: Causal modelling of dementia: AI-powered patient digital twins from multimodal observational data

Project reference: DT4H_07_2022

1st supervisor – Jorge Cardoso- School of Biomedical Engineering & Imaging Sciences

2nd supervisor – Seb Ourselin– School of Biomedical Engineering & Imaging Sciences

Aim of the project

This project aims to model disease progression and clinical interventions in patients presenting themselves with memory complaints, in the context of dementia. This will be achieved by learning the causal relationships and interactions between the clinical patient presentation, quantitative biomarkers extracted from blood samples and imaging data, and therapeutic/symptomatic interventions directly from observational data using advanced AI models. Through the use of causal graph discovery, the proposed model will be able to not only predict the most likely patient progression at the individual level, but also simulate how the disease would progress under different hypothetical clinical interventions via counterfactual modelling. By building on observational data (i.e. data from SLaM), the proposed causal AI models will provide insights not only about the expected and known interactions between clinical variables of interest (cognitive, blood, imaging), but also better understand the real-world behaviour of patients and what elements ultimately impact outcome.

Project description

Alzheimer's disease (AD), and dementia in general, is a key challenge for 21st-century healthcare. The statistics are sobering (Winblad et al., 2016): in a single year, 47 million people world-wide suffer from dementia, of which AD is the most common cause; dementia costs more than $800 billion worldwide, which is more than 1% of the aggregate global gross domestic product; AD might contribute to as many deaths as does heart disease or cancer. There are no available treatments that can cure or even slow the progression of AD – all clinical trials into putative treatments have failed to prove a disease-modifying effect. One key reason for these failures is the difficulty in identifying a group of patients at early stages of the disease, where treatments are most likely to be effective.

While early and accurate diagnosis of dementia can be challenging, clinical and mental examinations are often used in conjunction with quantitative biomarker measurements taken from medical imaging data (e.g. MRI, PET, etc) and cerebro-spinal fluid (CSF) samples extracted from lumbar puncture to diagnose. It has been shown (Jack Jr et al., 2010, 2013; Aisen et al., 2010; Frisoni et al., 2010) that clinically-observable changes and disease-related biomarkers of AD become abnormal/change at different stages and intervals even before symptom onset, suggesting that together they can be used for accurate prediction of onset, overall disease staging and expected progressions at the individual level. Blood-based biomarkers provides a less expensive and effective assessment for screening, staging, and monitoring of patients with neurodegenerative disorders. The ratio of two forms of plasma amyloid-ß peptide (Aß42/Aß40), isoforms of plasma phosphorylated tau (P-tau181, P-tau217, P-tau231), and plasma NfL are the most published blood biomarker related to neurodegenerative disorders. Although individually these plasma AD biomarkers have shown relatively fair accuracy for detecting pathology, they have limitations in accurate diagnosis of AD.

Recent studies (Palmqvist et al., 2021; Janelidze S et al., 2021) support the idea of using combinations of biomarkers with other information to improve diagnostic prediction.

Several approaches for modelling and predicting AD-related target variables (e.g. clinical diagnosis, cognitive, blood and imaging biomarkers, long-term progression, outcomes, etc) have been proposed, leveraging multimodal biomarker data available in AD clinical research studies. Traditional approaches model the relationships of the target variables with other known variables via predictive models, exploiting correlations and other non-linear relationships in the data. Examples include regression of the target variables against clinical diagnosis (Scahill et al., 2002), blood/plasma biomarkers (Beltran et al, 2020), cognitive test scores (Yang et al., 2011; Sabuncu et al., 2011), rate of cognitive decline (Doody et al., 2010), and retrospectively staging subjects by time to conversion between diagnoses (Guerrero et al., 2016). Another approach involves supervised machine learning techniques such as support vector machines, random forests, and artificial neural networks, which use pattern recognition to learn the relationship between the values of a set of predictors (plasma and imaging biomarkers) and their labels (diagnoses). These approaches have been used to discriminate AD patients from cognitively normal individuals (Kloppel et al., 2008; Zhang et al., 2011), and for discriminating at-risk individuals who convert to AD in a certain time frame from those who do not (Young et al., 2013; Mattila et al., 2011; Palmqvist et al., 2021; Janelidze S et al., 2021).

These approaches, however, only model correlational relationships and do not attempt to separate correlation from causation. By working on clinical research data with highly harmonised data collection, current models can achieve outstanding performance. Their performance, however, degrades significantly when used on real-world observational data. Observational data is often biased, noisy, polluted by external confounds, and missing-not-at-random, breaking many of the assumptions of current models. Causal modelling and inference allow one to disentangle these effects and better distinguish correlation from causation. Also, via counterfactual inference, one can simulate what would the outcome be under certain interventional scenarios. In short, causal inference from observational data provide us the tools to emulate a perfect hypothetical randomized trial: known as the target trial [Hernán et al. NEJM, 2021].