Objectives: Applying new machine learning and causal inference techniques for a better understanding of real-world drug effects

1: Compare machine learning and traditional methodologies in pharmacoepidemiology research ⟶ Promote the **robust** analysis of real-world data.

2: Gain a better understanding of real-world treatment effects of anti-diabetes therapeutics ⟶ Leverage real-world data to **better inform patients and clinicians.**

3: Study real-world heterogeneous treatment effects for anti-diabetes therapeutics ⟶ Develop evidence to **inform new hypotheses.**

Study design: Machine learning and causal inference methods will be applied to a baseline target trial emulation of an important cardiovascular outcome trial

Traditional methods

Base target trial emulation of the EMPA-REG RCT

The EMPA-REG trial was a seminal RCT that helped establish the cardioprotective effects of an SGLT2i, empagliflozin, in patients with diabetes and atherosclerotic CVD.

- Protocol for the EMPA-REG target trial - *ethical approval completed and protocol publication pending December 2023.*
- Descriptive study of SGLT2i prescribing in UK clinical practice. *[Work stream 1]*
- Target trial emulation to determine real-world treatment effects. *[MSc project]*

Medical Statistics MSc work

Advanced methods

Application of advanced covariate adjustment techniques

- Validate NLP methods to derive symptom-based covariates from free text notes. *[Work stream 2]*
- Descriptive study of NLP-derived symptom covariates. *[Work stream 2]*
- Compare various ML methods for covariate adjustment (HDPS, NLP-PS, ML-PS) in a target trial framework. *[Work stream 3]*

Determination of real-world and heterogeneous treatment effects

- Heterogeneous treatment effects within the target trial and comparison with subgroup analysis in clinical trials. *[Work stream 4]*
- Application of HTE-ML methods to identify subgroups where therapeutics might have differential effects. *[Work stream 3, 4]*
- Submission of results to NICE or other regulatory bodies and consideration of new hypothesis for confirmatory clinical trials.

Proposed PhD Fellowship

PhD Impact:

Patients and clinicians:
What is the real effect of taking this drug?

Drug regulators, guideline committees and pharmaceutical industry:
- *What gaps in evidence can this data address?*
- *What new hypotheses can be generated from real-world evidence?*

Statisticians and Epidemiologists:
- *What is the optimal way to analyse real-world clinical data?*
- *How can ML techniques reduce residual confounding?*

Timeline:

Year 1: Preparation
- Publication of TTE [work stream 1]
- Validate NLP methods [work stream 2]
- Descriptive NLP work [work stream 2]

Year 2: Propensity score
- NLP-PS
- HDPS
- ML-PS

Year 3: PPI and treatment effect heterogeneity
- PPI panels
- HTE work - causal forests and meta-learners

Figure 1: Hypotheses and study design for proposed PhD. SGLT2i: sodium-glucose cotransporter-2 inhibitor; RCT: randomised controlled trial; RW: real-world; NLP: natural language processing; TTE: target trial emulation; ML: machine learning; HTE: heterogeneous treatment effects; NICE: National institute for health and care excellence; PS: propensity score; HDPS: high-dimensionality PS; PPI: patient and public involvement.