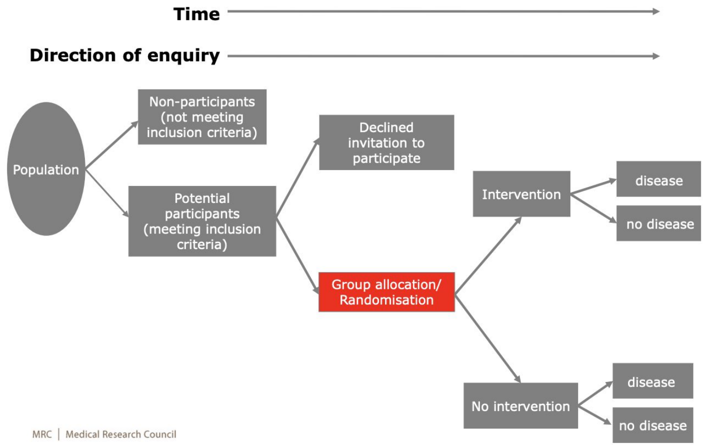
Statistical inference in emulated trials using observational data

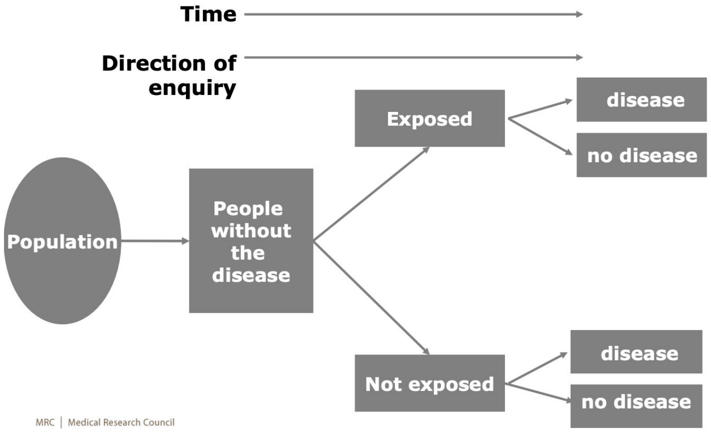
Applying for funding presentation

Juliette Limozin
Supervised by Dr Li Su, MRC Biostatistics Unit

Introduction

- Randomised trials are the gold standard for causal inference estimation, but they are often *expensive, time consuming and not suitable due to ethical constraints*
- Observational data is largely available and a good alternative to estimating cause and effect, but patients are not randomly allocated to treatment or control so differences in the observed effect might be partially or fully *due to the differences in the individuals* rather than the differences in treatment.


Time
Direction of enquiry
Population
Non-participants (not meeting inclusion criteria)
Potential participants (meeting inclusion criteria)
Declined invitation to participate
Group allocation/ Randomisation
Intervention
No intervention
disease
no disease
disease
no disease
MRC | Medical Research Council



Research question

- An emulated target trial method has been proposed by Herman and Robins (2016), which selects individuals that meet the eligibility criterias of a 'target' randomised trial in large observational data to compare the outcomes between treated and untreated

This dissertation aims to obtain **statistical inference** (such as calculating confidence intervals) of treatment effects or survival outcome in emulated target trials

Aims

After validation on simulated data, the project would benefit from testing the statistical inference methods on:

- **Large open access** observational data (e.g. UK Biobank) for emulated target trials on common diseases
- **Data from large pharmaceuticals** (e.g. Roche) for target trials on the effect of their products on patients

The findings would be incorporated in a **R package** currently under development to facilitate target trial emulation.

Data sources and analysis

Simulated data in a clinical context will be used to evaluate the statistical methods.

Advantages of simulated data:

- Distributions of the treatments and effects are know, so true values of statistics drawn from the distributions can be compared to the estimated values
- They are modifiable, allowing for validation in various settings

Disadvantages:

- Simulated data is a *simulation*, not real observed data

Discussion

Strengths:

- Simulated data allows for a *controlled setting* in which the validation of statistical inference methods is reliable

Limitations:

- Statistical inference on simulated data can be *computationally expensive and time-consuming*; the project will most likely require access to high-performance computing servers

Timeline

February - Mid March

Getting familiar with the R package, set up access to HPC servers

Literature review on the existing statistical inference methods

Theoretical work on improving the existing methods or building new ones

Mid March - June

Validate the methods through simulated data

July

Final write up of results and discussions