**3.4 Machine Learning meets Biostatistics II**

Qingyuan Zaho

<http://www.statslab.cam.ac.uk/~qz280/teaching/>

<http://www.dagitty.net/>

<http://www.statslab.cam.ac.uk/~qz280/talk/ccaim-summer-school-2022/>.

The 16-hour course is given to Part III mathematicians/statisticians and the course webpage is <http://www.statslab.cam.ac.uk/~qz280/teaching/causal-2021/>.

**Randomization and potential outcomes**

* Motivating examples: Vitamin studies
  + 1990s, several studies found strong inverse association
  + Conducted RCT showed that supplementation with antioxidants does not protect against these diseases
  + <https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(04)16260-0/fulltext>
* What went wrong?
  + Confounder = Common cause of treatment and effect
    - Cannot be removed completely. Always potential issue with unmeasured confounders
  + How can we balance observed confounders? Better design (e.g. blocking)
  + Randomisation
* Randomisation as a basis of inference
  + Randomisation now regarded as gold standard for causal inference – was difficult to accept
    - Example:
      * Physician allowed to administer a promising new drug to 5/10 patients
      * Physician thinks the best way to prove effectiveness of the drug is to give it to the 5 patients that they think are the most ill
      * Flaw
    - Randomization introduces an objective basis of inference which anyone else can use

**DAG models:**

* Conditional independence
  + Two ways of testing:
    - Conversion to undirected graph
      * Moralisation
    - D-separation on directed graphs
  + Both criteria are mathematically equivalent
    - Factorisation according to DAG

**Causal DAGs**

* Correlation is not causation
  + Model may not generalise to other settings

## **Readings**

The following books/articles are optional. I am providing a short (personal) verdict to help you navigate the literature.

* [Causal Inference for Statistics, Social, and Biomedical Sciences](https://idiscover.lib.cam.ac.uk/permalink/f/t9gok8/44CAM_ALMA51529379970003606) by Guido Imbens and Donald Rubin [IR]. This book provides a gentle introduction to potential outcomes and statistical methods for simple randomised experiments and observational studies with no unmeasured confounders.
* [Causal Inference: What If](https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/) by Miguel Hernán and James Robins [HR]. This book provides a comprehensive treatment for causal inference without and with models.
* [Causality: Models, Reasoning, and Inference](https://idiscover.lib.cam.ac.uk/permalink/f/1ii55o6/44CAM_ALMA51527615290003606) by Judea Pearl [Pearl]. A great book if you are interested in the philosophical debates in causal inference.
* [Statistical Models: Theory and Practice](https://idiscover.lib.cam.ac.uk/permalink/f/1ii55o6/44CAM_ALMA51527574630003606) by David Freedman. A less technical textbook is well suited for someone who wants to learn the basic ideas in causal inference through practical examples.
* [Graphical Models](http://www.statslab.cam.ac.uk/~qz280/teaching/causal-2019/reading/Lauritzen_1996_Graphical_Models.pdf) by Steffen Lauritzen. A good reference for probabilistic graphical models.
* [Observational Studies](http://www.statslab.cam.ac.uk/~qz280/teaching/causal-2019/reading/Rosenbaum_2002_Observational_Studies.pdf) by Paul Rosenbaum. A good book for randomisation inference and sensitivity analysis.
* [Mostly Harmless Econometrics: An Empiricist’s Companion](https://idiscover.lib.cam.ac.uk/permalink/f/t9gok8/44CAM_ALMA51582506410003606) by Joshua Angrist and Jörn-Steffen Pischke. Very clearly written book from an applied econometrics point of view, with a lot of useful intuitions.