Session2:
An introduction to statistical thinking & data workflows

Dr Nicola Foster
UCL Respiratory
nicola.foster@ucl.ac.uk

GitHub: nicolacfoster

Outline of this session

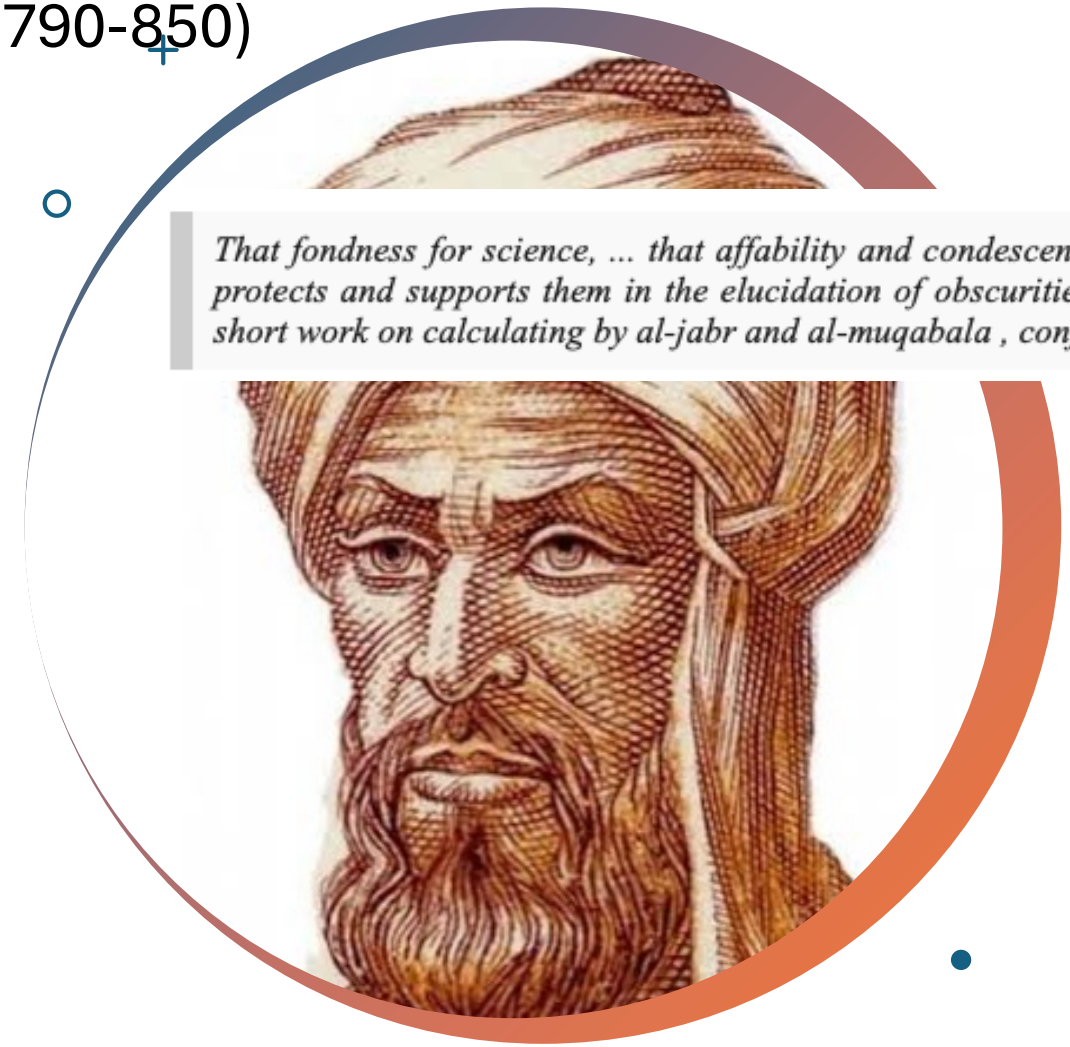
- What is statistics?
- Some famous statisticians
- Statistical thinking in relation to research or knowledge generation*
- How do our approach to knowledge generation influence how we analyse and interpret the data?
- Data processing and data workflows

Do you learn from the data, or
do you first outline your previous knowledge &
assumptions?

What is statistics?

- The science of the collection, organization, analysis, interpretation and presentation of data.
- The practice of drawing inferences from a representative sample about the whole population.
- Develop models (simplified mathematical rules/ representations of the real world) to make decisions.

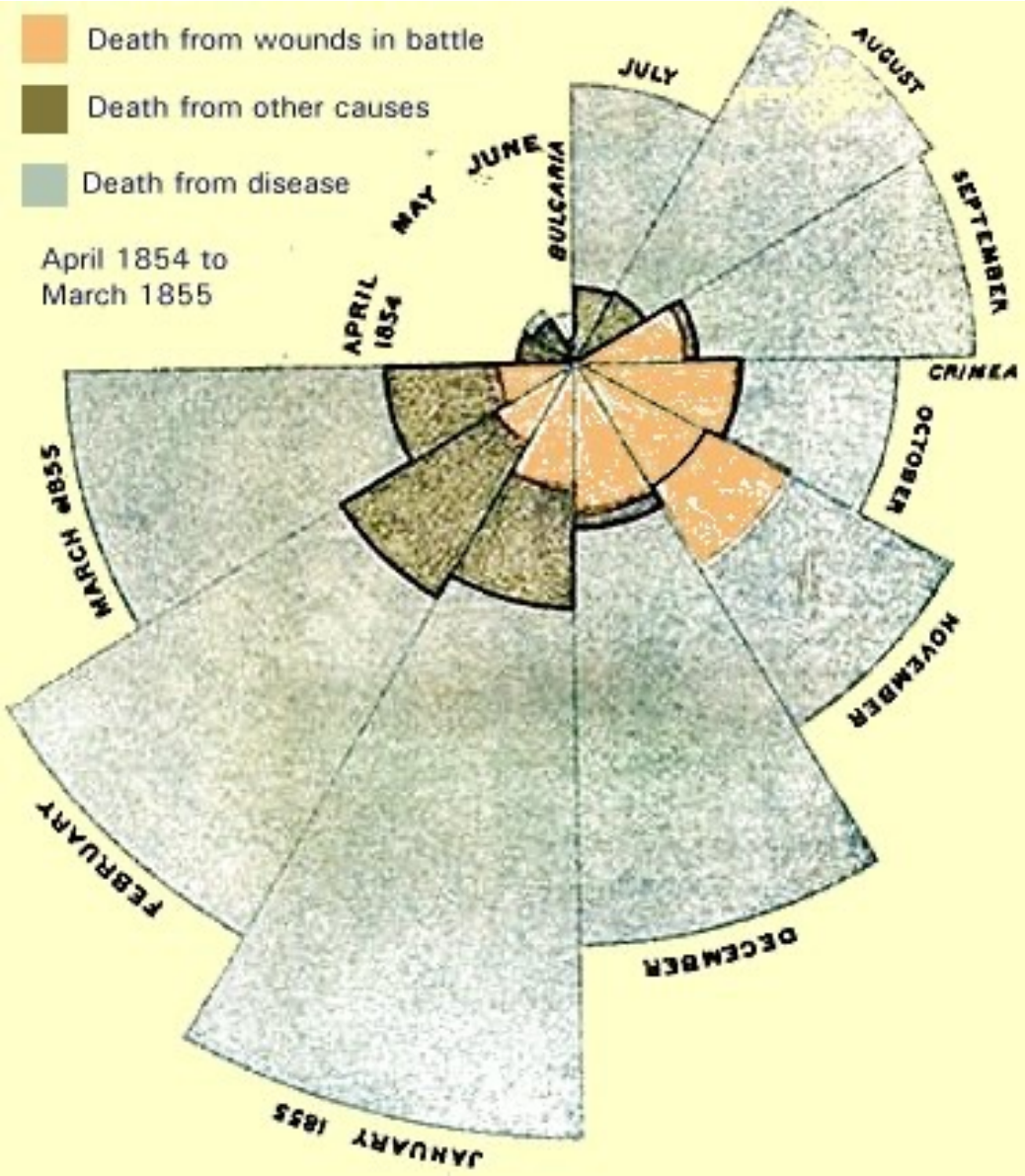
Al-Khwarizmi (790-850)



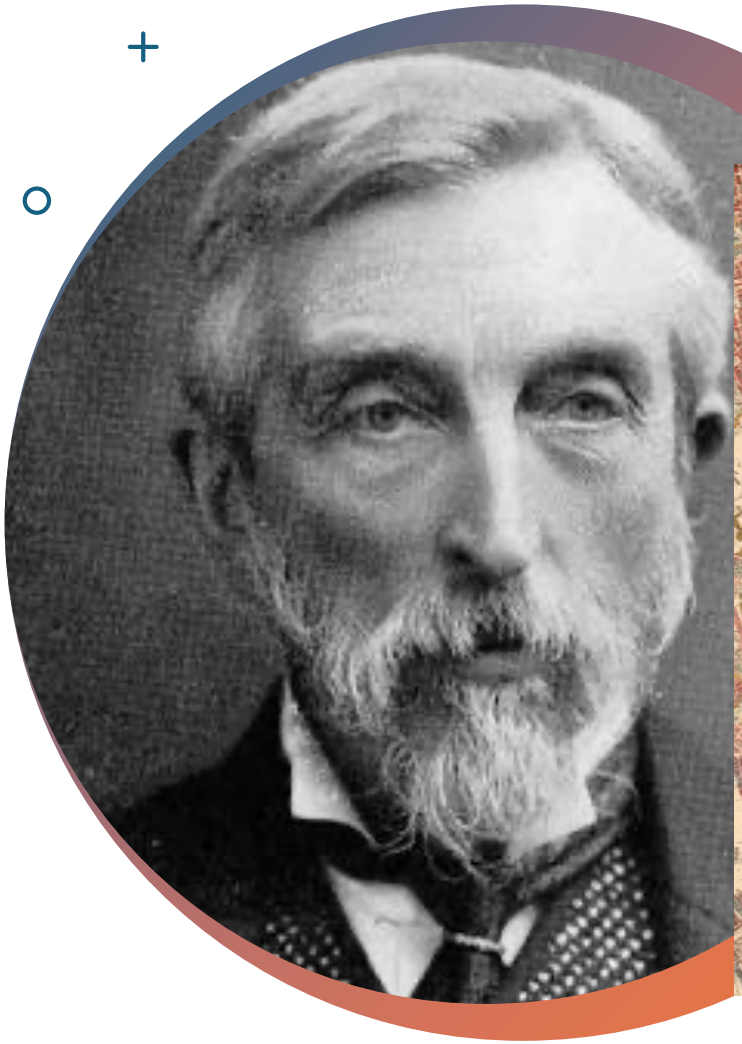
> *That fondness for science, ... that affability and condescension which God shows to the learned, that promptitude with which he protects and supports them in the elucidation of obscurities and in the removal of difficulties, has encouraged me to compose a short work on calculating by al-jabr and al-muqabala , confining it to what is easiest and most useful in arithmetic.*

- The inventor of algebra
- Introduced Hindu-Arabic numerals to Europe
- Developed the concept of zero
- He mostly used words rather than letters to solve equations
- Application of his work at the time was solutions for land distribution, inheritance and salary distribution.

Florence Nightingale (1820-1910)



Charles Booth (1840-1916)



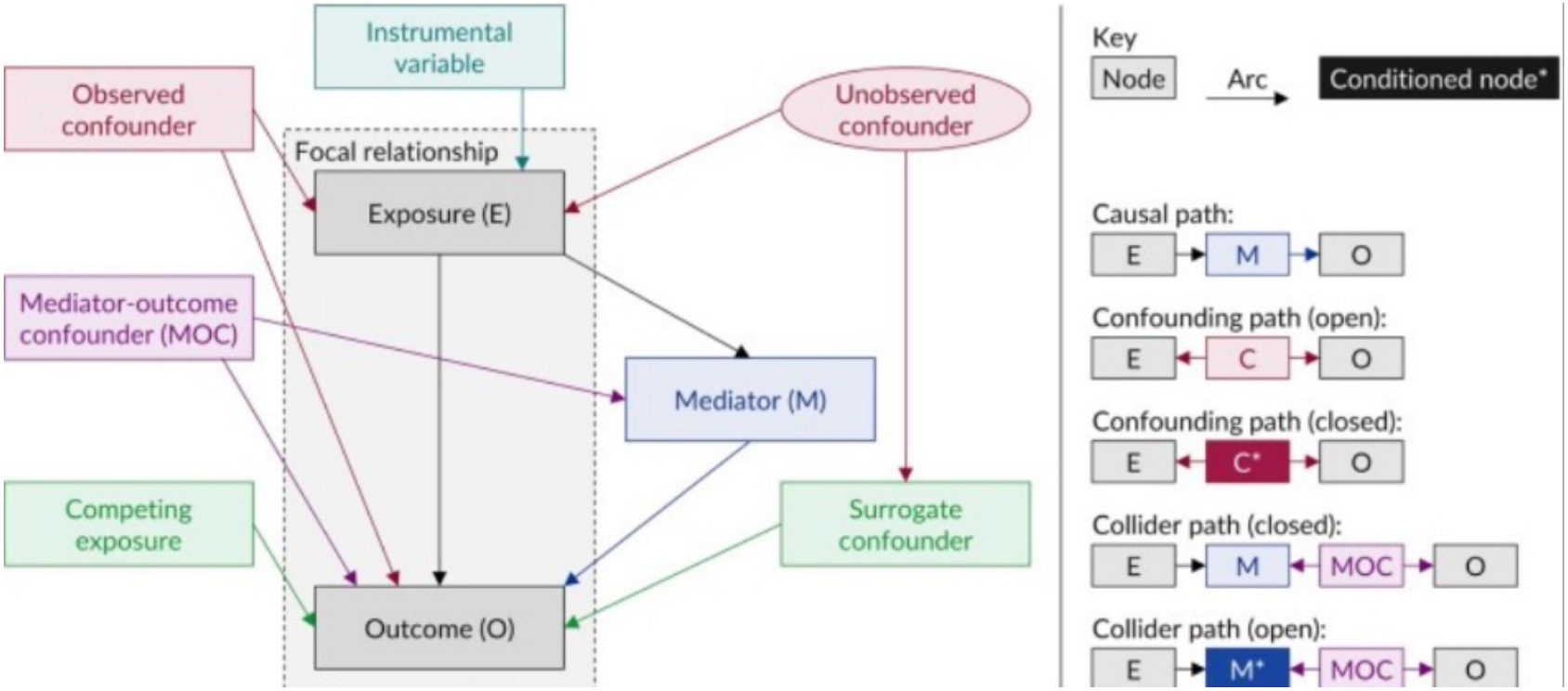
Karl Pearson (1857-1936)

- Founder of the world's first statistics depart at UCL in 1911
- Very diverse education including science, law and philosophy
- Biometrics – developed the chi-squared test, standard deviation, correlation and regression coefficients, p-value
- Method of moments for fitting distributions to samples
- Contributed to foundation of statistical hypothesis testing theory and statistical decision theory
- Principal components analysis
- Controversial because proponent of social Darwinism and eugenics (scientific racism)

Statistical thinking in relation to knowledge generation

- Do you learn from your data, or do you first outline your previous knowledge & assumptions then test against the data?
- We want to reduce subjectivity and be as rational as possible, approaches include:
  - Learning from the data – machine learning algorithms/ AI. But needs guidance and how is this decided? Do we have a process/ framework for this?
  - Or setting up randomized experiments such as individually randomized clinical trials or cluster randomized trials. Strict rules about how these are conducted.

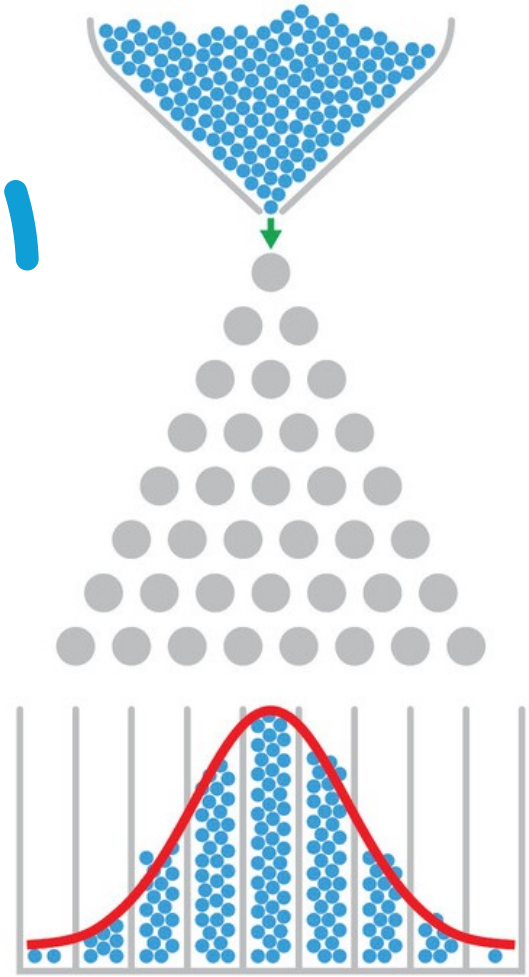
Directed acyclic graphs



Tennant et al. Use of DAGs to identify confounders in applied health research: review and recommendations. PMID: 33330936.

Frequentist statistics

- Based on probability theory
- Work towards drawing conclusions using data from a sample of people about a population using the frequency/ proportion of findings in the data
  - Underlying this are theories such as Galton's theory which is that under normal conditions, the distribution of variation is gaussian (normal) – central limit theory
  - Exceptions to this theory, certain types of data not expected to be normally distributed – non-parametric tests



Bayesian statistics

- Based on Bayes understanding of probability
- Conditional probability
  - The probability of event A given event B
  - An example of this is in how we understand sensitivity and specificity, we include a chance that we may be wrong – false positive and false negatives.
- In application, we set up a prior then calculate posterior probabilities based on a prior and likelihood
  - In other words - the prior probabilities are updated through an **iterative** process of data collection/ assessment
  - The strength of this approach to statistics and estimating effect sizes, is how uncertainty is assessed and represented

How do our approach to knowledge generation influence how we analyse and interpret the data?

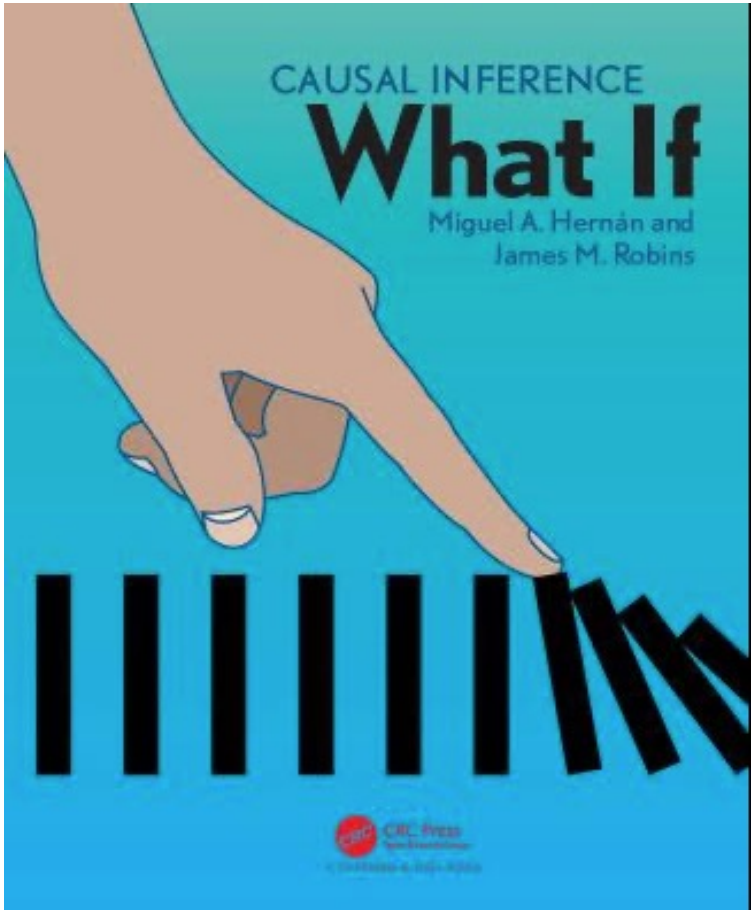
- Focus on effect sizes and uncertainty rather than p-values and "statistical significance"
  - Avoid arbitrary cut-off points –p-values
  - Could lead to incorrect conclusions especially if not interpreted within the context of the whole study
- Alternatives
  - Effect size and confidence interval

    Interpreted through study design
  - Bayes factor
  - Akaike information Criterion

McElreath. Statistical rethinking (book & lectures on youtube)

Causal inference

- A scientific method
- The process of determining the (independent) effect of a particular phenomenon/ intervention that is a component of a larger system
- A set of methods for examine and quantify the actual relationship in the data we are analysing.



Data workflows

Data workflows

- Approaching your data analysis in a systematic way, that can be documented and audited.
- Steps could include:
  1. Data analysis plan, informed by your research question and the existing literature
  2. Data validation
  3. Create required variables for your analysis
  4. Write analysis script/ code
  5. Implement analysis and discuss results with investigators
  6. Revise and finalise
- Currently, I use R and Quarto for these tasks. But could include any programming language (script based).

Step1: Data validation

- Review all data documentation
  - Data dictionaries
  - Review variables and discuss variable definitions with those who collected the data.
  - Plot all data distributions.
  - Look closely at minimum and maximum values. Identify any biologically or otherwise implausible values – set these to missing
- Keep the original dataset provided, your validation code, and a record of any changes you made to the dataset before analysis

Step2: Generate variables required for your analysis

- Perform calculations to generate new measures
- Or create an index
- Agreement with collaborators on how these measures should be created
- Remember to label these generated variables, so you know how they were generated

Step3: Write (a first draft) of your analysis script

- Review whether the statistical models used are appropriate give the data distributions (from step1)
  - Be clear on underlying assumptions of statistical models
  - Use formal statistical tests where appropriate
- Descriptive analyses
- Primary outcome, secondary outcomes, the explanatory analyses and the sensitivity analyses

Step4: Discuss initial results with other investigators

- Consider using Quarto to document initial results
  - A script-based notebook that allows for the visualization of code plus results from analyses
  - Output can be a word document or html
  - Can use multiple programming languages in the same notebook
- Test and run your analysis or scripts separately first, then use summaries/ graphical visualisations in your Quarto document.

```yaml
---
title: "Analysis Report"
author: "Produced by Nicola Foster, UCL"
format: docx
editor: visual
---
```

Evaluating FEV1Q as a race-neutral assessment of lung function

This document details the reproduction of the analysis for the paper led by Ayadh Alayadhi. This version is using the revised GECO analysis dataset (see e-mail trial from 27 September 2024, using the dataset provided by Prof Julie Barber). Changes made to this dataset involved constructing a socio-economic position (SEP) variable, updating some of the clinical variables to exclude any outlier values by setting biologically implausible values to missing - and to match the paper using the same variables written by William Checkley (see datareport.qmd).

```r
{r}
#| label: load
#| include: false
#| warning: false
#| echo: false

library(haven)
library(ggplot2)

## CHECK THE DATASET AND THAT VALUES MATCH UP - BASED ON DISCUSSION WITH JULIE BARBER
data1 <- read_dta("/Users/ucl/Library/Mobile Documents/com~apple~CloudDocs/project_GECO/data/geco5.dta")
# View(data1)
```

Clinical and demographic characteristics of study participants by site.

```r
{r}
#| label: demographics
#| echo: false
#| fig-cap: Table1. Baseline characteristics of study participants.

# loads packages
library(gtsummary)
library(labelled)
library(expss)

# preparing the variables
country <- as.factor(data1$site)
age <- as.numeric(data1$calculatedage)
gender <- as.factor(data1$sex_cat)
packyears <- as.numeric(data1$packyears)
bmi_cat <- as.factor(data1$bmi_cat)
biomass <- as.factor(data1$biomass)
copdmeasured <- as.factor(data1$copdm)
obstructed <- as.factor(data1$obstructed)
sep <- as.factor(data1$SEP_1)
```

Step5: Revise and finalise

- For a completed analysis – folder with the following files:
  1. Data analysis plan*
  2. Original dataset provided
  3. Data dictionary*
  4. Analysis dataset – after you have generated the variables used in your analysis*
  5. Code: data validation
  6. Code: generating variables for analysis
  7. Code: analysis code, includes sensitivity analyses and formal model assessments*
  8. Code notebook to share with collaborators
- Also to keep a record of methodological comparisons done or decisions made.
- Not all of this is needed for sharing with collaborators but important to keep.

Step5: Revise and finalise

- For a completed analysis – folder with the following files:
  1. **Data analysis plan**
  2. Original dataset provided
  3. **Data dictionary**
  4. **Analysis dataset – after you have generated the variables used in your analysis**
  5. Code: data validation
  6. Code: generating variables for analysis
  7. **Code: analysis code, includes sensitivity analyses and formal model assessments**
  8. Code notebook to share with collaborators
- Also to keep a record of methodological comparisons done or decisions made.
- Not all of this is needed for sharing with collaborators but important to keep.

Step6: Open Access

- Increasingly expected to make code and data files available with publication of papers
- An output on it's own – get a doi and some of these 'products' can be used/ adapted by others and cited
- But some implications to consider
  - Permission from those who provided data (data protection office)
  - Review informed consent, and the study protocol
  - Ethics approval, especially if linked datasets
  - May depend on the funder – be sure to check.
- What do you share and how?

Open access lifecycle for a project

Publishing your study protocol/ data analysis plan

BMC Part of Springer Nature Search

Trials

Home About Articles In Review Submission Guidelines News COVID-19 Submit manuscript

Study protocol | Open access | Published: 24 April 2023

Evaluating the equity impact and cost-effectiveness of digital adherence technologies with differentiated care to support tuberculosis treatment adherence in Ethiopia: protocol and analysis plan for the health economics component of a cluster randomised trial

Nicola Foster, Amare W. Tadesse, Christopher Finn McQuaid, Lara Gosce, Tofik Abdurhman, Demelash Assefa, Ahmed Bedru, Rein M. G. J. Houben, Kristian van Kalmthout, Taye Letta, Zemedu Mohammed, Job van Rest, Demekech G. Umeta, Gedion T. Weldemichael, Hiwot Yazew, Degu Jerene, Matthew Quaife & Katherine L. Fielding

Trials **24**, Article number: 292 (2023) | Cite this article

Register your study – systematic review

NIHR | National Institute for Health and Care Research

PROSPERO

International prospective register of systematic reviews

Print | PDF

Understanding implementation of digital adherence technologies for active and latent tuberculosis: A systematic review using the RE-AIM framework

Shruti Bahukudumbi, Chimweta Chilala, Ramnath Subbaraman, Kevin Schwartzman, Katherine Fielding, Mona Salaheldin Mohamed, Miranda Zary, Cedric Kafie

Register your study – trial/ clinical study

ISRCTN registry

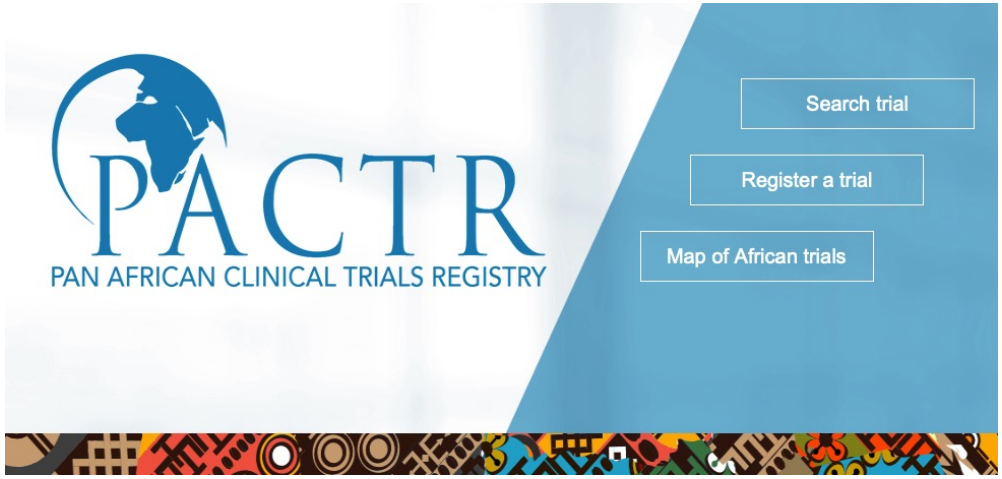
The ISRCTN registry is a primary clinical study registry recognised by the World Health Organisation (WHO) and the International Committee of Medical Journal Editors (ICMJE) that accepts all clinical research studies (whether proposed, ongoing or completed), providing content validation and curation and the unique identification number necessary for publication. All study records in the database are freely accessible and searchable.

ISRCTN supports transparency in clinical research, helps reduce selective reporting of results and ensures an unbiased and complete evidence base.

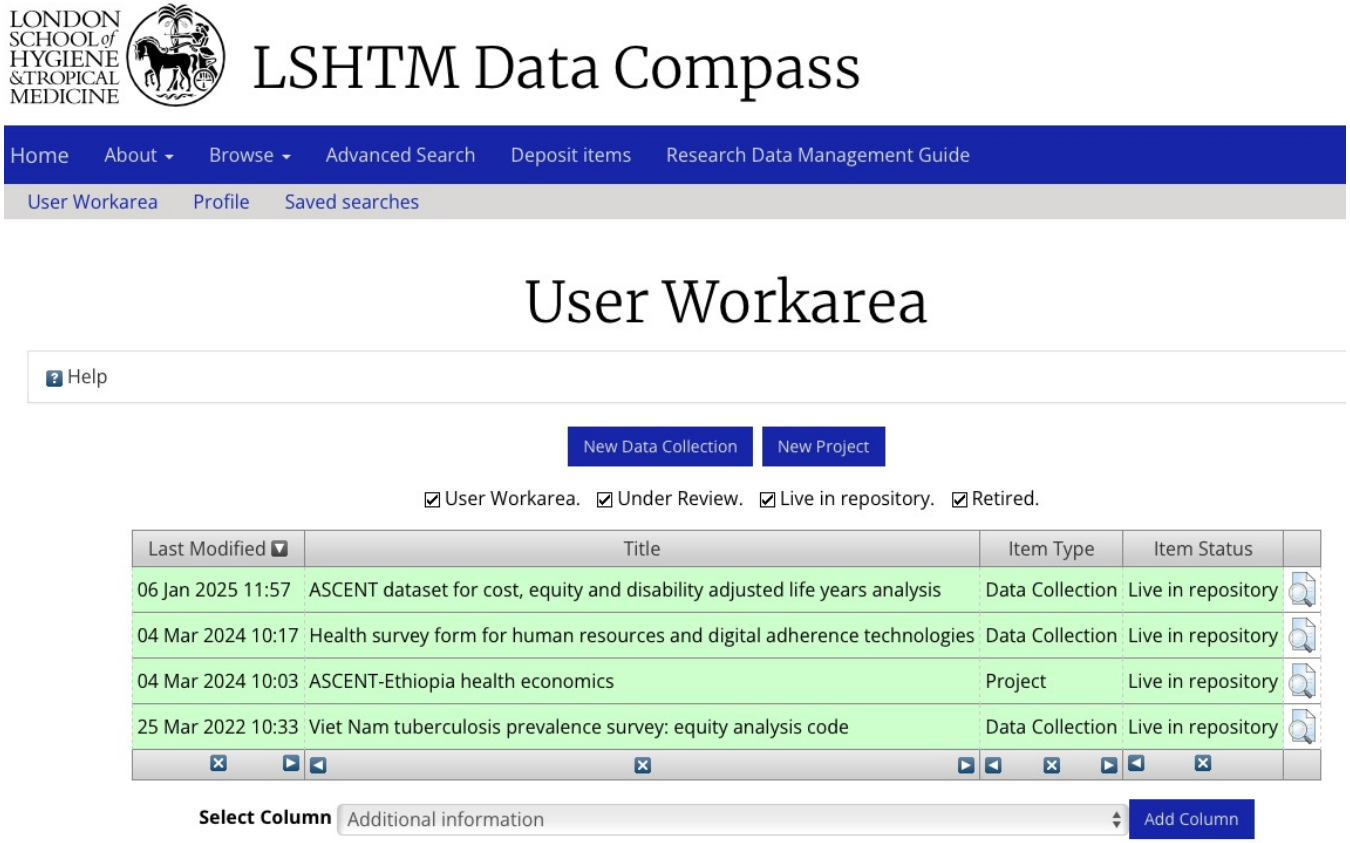
The registry aims to include all interventional and non-interventional clinical studies that prospectively involve UK participants and evaluate biomedical or health-related outcomes.

Studies conducted outside the UK or considered to be non-clinical studies (e.g. public health studies) can be registered on ISRCTN.

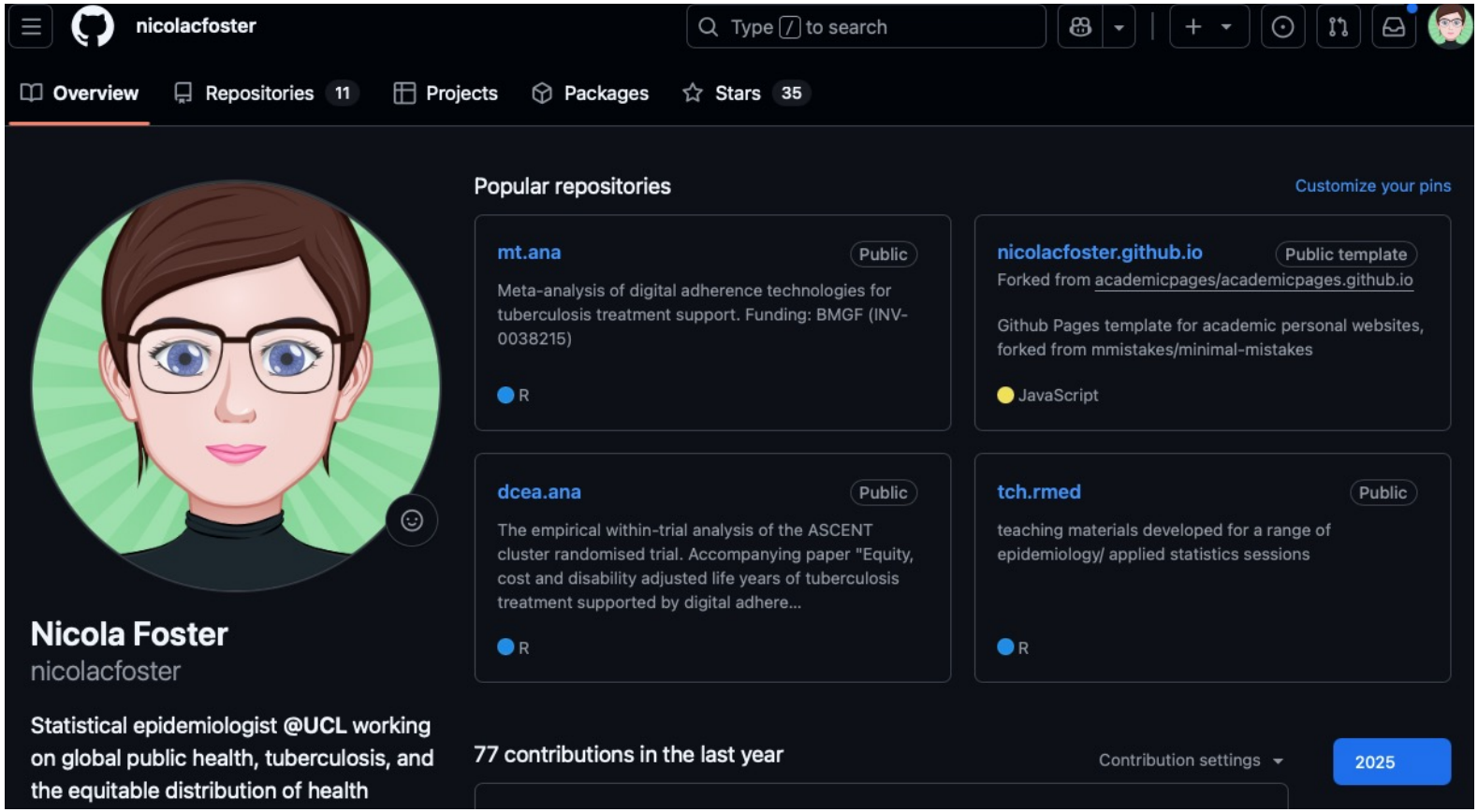
Studies should ideally be registered prospectively (before recruitment starts). ISRCTN also accepts studies registered retrospectively once they are underway or after completion.



Sharing data – long-term data repositories



Sharing code – GitHub and Zenodo



nicolacfoster / dcea.ana

Type / to search

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

dcea.ana Public

Pin | Unwatch 1 | Fork 0 | Star 0

main | 1 Branch | 1 Tag | Go to file | t | Code

nicolacfoster uploads the imputation and dcea models' code 4dd9022 · 2 months ago 4 Commits

LICENSE	Initial commit	3 months ago
README.md	Update README.md	3 months ago
imp3.R	uploads the imputation and dcea models' code	2 months ago

README | CC0-1.0 license

dcea.ana

The empirical within-trial analysis of the ASCENT cluster randomised trial. Funding: UNITAID (2019-33-ASCENT). Paper: "Equity, cost and disability adjusted life years of tuberculosis treatment supported by digital adherence technologies and differentiated care in Ethiopia: a trial-based distributional cost-effectiveness analysis". doi: 10.1101/2024.07.28.24310767. link: https://www.medrxiv.org/content/10.1101/2024.07.28.24310767v2.

About

The empirical within-trial analysis of the ASCENT cluster randomised trial. Accompanying paper "Equity, cost and disability adjusted life years of tuberculosis treatment supported by digital adherence technologies and differentiated care in Ethiopia: a trial-based distributional cost-effectiveness analysis".Funding: UNITAID (2019-33-ASCENT)

Readme
CC0-1.0 license
Activity
0 stars
1 watching
0 forks

Releases 1

ASCENT DCEA Latest
on Dec 9, 2024

Packages

No packages published
Publish your first package

Languages

R 100.0%

Questions?

- If you have topics that you would like me to discuss in detail/ show examples of, you can log them here:

https://forms.gle/8S3GFotKadZDbRfa7