Foundations in Causal Thinking for Health Data Statisticians

Erica E. M. Moodie

McGill University, Biostatistics
erica.moodie@mcgill.ca
www.ericamoodie.com

STRATOS INITIATIVE

On behalf of TG7

> everything that becomes or changes must do so owing to some cause; for nothing can come to be without a cause
>
> *Plato*

- "Epidemiological research is, almost exclusively, concerned with *etiology* of illness", where etiology = causal origin of illness (Miettinen & Karp, 2012).
- In fact, the goal of most statistical analyses is to uncover causal relationships.

Causality

- There is no agreement on the definition of causality, particularly across disciplines (or across centuries!).
- In 1890, Koch proposed criteria to establish a 'causative relationship' between a microbe and a disease; imperfect but reasonable - but only for pathogens.
- Pearl (2009, p. 25-26), a computer scientist and a leader in the field of modern causality, does not explicitly define causality at all, refers to causal relationships as "stable" and "ontological".
- Meinshausen, Peters & Buehlmann (2016) similarly deem causal relationships to be present when multiple data sources produce 'invariant prediction'.

Bradford Hill

- Earliest, and best known, ideas in epidemiology on causality are the (non-)criteria given by Sir Austin Bradford-Hill in 1965:
    1. Strength
    2. Consistency
    3. Specificity
    4. Temporality
    5. Biological gradient
    6. Plausibility
    7. Coherence
    8. Experiment
    9. Analogy
- A group of conditions to *assess* (not establish) causality.

Bradford Hill

- Less well-known is Bradford Hill's wide-ranging lecture on 'The Statistician in Medicine', recently reprinted in *Statistics in Medicine* in celebration of 40 years since its inception.
- Three themes:
  - knowledge of the area of application,
  - types of data to provide evidence and how it is gathered (including the poor experiment that is 'nature'),
  - drawing conclusions from evidence.
- Still very relevant!

‘The Statistician in Medicine’: On medicine

> [The statistician] must learn a great deal of medicine and [...] not only have facility in speaking two languages, he must be able to think in two.

‘The Statistician in Medicine’: On medicine

> [The statistician] must learn a great deal of medicine and [...] not only have facility in speaking two languages, he must be able to think in two.

- Bilingualism!
- Context is critical implementing any analysis
- Subject-matter knowledge must be used to inform modelling, but equally critical to note what is known vs. what is assumed, and whether the data themselves were collected fairly.
- DAG

‘The Statistician in Medicine’: On medicine

204 *N. Röhrig et al. / Journal of Clinical Epidemiology 67 (2014) 199–206*

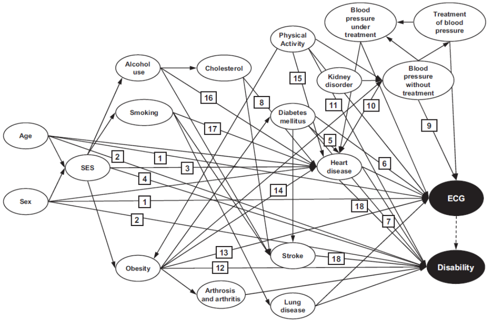


Fig. 1. DAG derived from literature and expert knowledge. Nodes represent variables and arrows represent causal associations. Dark-colored nodes label ECG findings and disability, representing exposure and outcome, respectively. Pale-colored nodes represent possible confounding factors. Numbers represent available information from the literature (see Table 1 at www.jclinepi.com for full references). SES, socioeconomic status; ECG, electrocardiography.

‘The Statistician in Medicine’: On causality

- Among BH’s items, ‘strength of association’ is one that is not so easily dismissed.

- Among BH’s items, ‘strength of association’ is one that is not so easily dismissed.

> The question, on the other hand, may well be asked, what does one accept as overwhelming? When does a heap really become a heap? The answer, I submit, is not to be found tidily tucked up in the formulae of tests of significance, useful as they may be. In it there must always be an element of the subjective - the subjective judgment of the particular respondent, of you or me.

‘The Statistician in Medicine’: On causality

- Among BH’s items, ‘strength of association’ is one that is not so easily dismissed.

The question, on the other hand, may well be asked, what does one accept as overwhelming? When does a heap really become a heap? The answer, I submit, is not to be found tidily tucked up in the formulae of tests of significance, useful as they may be. In it there must always be an element of the subjective - the subjective judgment of the particular respondent, of you or me.

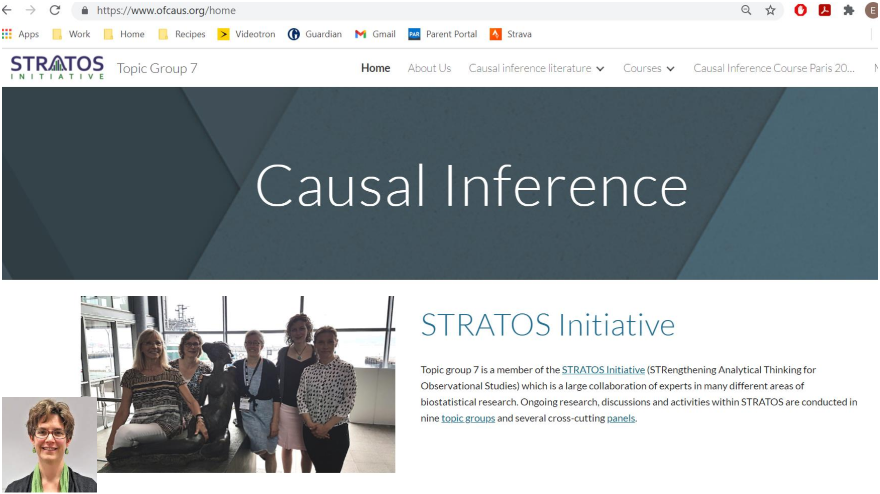
- ...so correlation *may be* causation. (Just be careful!)

‘The Statistician in Medicine’: On causality

We are continuously brought back to the fundamental question – what alternative explanation will fit a set of observations, what other differences between our contrasted groups could equally, or better, account for the observed incidences.

- This is where the ‘causal’ methods come in: we seek methods that reduce or eliminate alternative explanations such as *imbalance* in covariates between treatment groups (confounding, selection, missing data, etc.)

The STRATOS in Medicine: Causal inference!



The STRATOS in Medicine: Causal inference!

Received: 12 May 2019 | Revised: 10 May 2020 | Accepted: 5 August 2020
DOI: 10.1002/sim.8741

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

Check for updates

Formulating causal questions and principled statistical answers

Els Goetghebeur[1,2] | Saskia le Cessie[3] | Bianca De Stavola[4] | Erica EM Moodie[5] | Ingeborg Waernbaum[6] | "on behalf of" the topic group Causal Inference (TG7) of the STRATOS initiative

[1]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

[2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[3]Department of Clinical Epidemiology/Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Although review papers on causal inference methods are now available, there is a lack of introductory overviews on *what* they can render and on the guiding criteria for choosing one particular method. This tutorial gives an overview in situations where an exposure of interest is set at a chosen baseline ("point exposure") and the target outcome arises at a later time point. We first phrase relevant causal questions and make a case for being specific about the possible exposure levels involved and the populations for which the question is relevant. Using the potential outcomes framework, we describe principled definitions

The STRATOS in Medicine: Causal inference!

Tutorial: Formulating causal questions and principled statistical answers

On this page we include a link to our tutorial: Formulating causal questions and principled statistical answers, along with a data set and accompanying code. The purpose of the data is to illustrate concepts and estimation approaches by simulating a case study inspired by the Promotion of Breastfeeding Intervention Trial (PROBIT) a large randomized study in which mother-infants pairs across 31 Belarusian maternity hospitals were assigned either standard care or the possibility to follow a breastfeeding encouragement programme (Kramer MS et al. 2001). The aim was to investigate the effect of the programme and breastfeeding on a child's later development. In our simulation we go beyond generating the 'observed data' by also simulating for every unit in the study how different exposures would lead to different potential outcomes. The data set is called the simulation learner PROBITsim.

References

Kramer MS, Chalmers B, Hodnett ED, et al. Promotion of breastfeeding intervention trial (PROBIT) - A randomized trial in

the Republic of Belarus. Journal of the American Medical Association. 2001;285(4):413-420.

TITLE	LAST MODIFIED
Analysis with SAS A2.sas	5/16/19 Ingeborg Waernbaum
Analysis with SAS A3_A0.sas	5/16/19 Ingeborg Waernbaum
Analysis with SAS A3_A1.sas	5/16/19 Ingeborg Waernbaum
Analysis_code.R	5/16/19 Ingeborg Waernbaum
Description of data analysis.pdf	5/16/19 Ingeborg Waernbaum
Description of the data generating process.pdf	5/16/19 Ingeborg Waernbaum
Generate_simulation_learner.R	5/16/19 Ingeborg Waernbaum
PROBITsim2018_v12.dta	11/17/19 Ingeborg Waernbaum

The STRATOS in Medicine: Causal inference!

Tutorial: Formulating causal questions and principled statistical answers

...for survival outcomes (in progress)

...for time-varying exposures (future work)

On this page we include a link to our tutorial: Formulating causal questions and principled statistical answers, along with a data set and accompanying code. The purpose of the

Summary

- Causal estimation can be accomplished *by design*, but often requires the additional help of *modelling* to account for unplanned imbalances.
- The statistical framework for causal inference tries to formalize assumptions and make them as clear and explicit as possible; we mustn't forget that they are present and the foundation for our conclusions.

Summary

- Causal estimation can be accomplished *by design*, but often requires the additional help of *modelling* to account for unplanned imbalances.
- The statistical framework for causal inference tries to formalize assumptions and make them as clear and explicit as possible; we mustn't forget that they are present and the foundation for our conclusions.
- The stakes in medicine and public policy are high – it is worth investing energy and care to ensure our heap of evidence is convincingly high.