EPI289: Epidemiologic Methods III
Models for Causal Inference

INSTRUCTORS: Barbra Dickerman, Joy Shi

TEACHING FELLOWS: Lawson Ung (Head TF), Motohiko Adomi, Ruchita Balasubramanian, Soroush Moallef, Sakurako Okuzono, Dougie Zubizarreta

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Causal inference: a central task of science

- To estimate the causal effect of a treatment/exposure on an outcome

- Physics, chemistry, biology…
  - Experiments and observations
- Epidemiology, economics, sociology…
  - Mostly observations, some randomized experiments

Not all scientific questions are causal questions

- Not even all important questions are causal questions
- Scientists use data to ask three types of questions
  - Including health data scientists such as epidemiologists

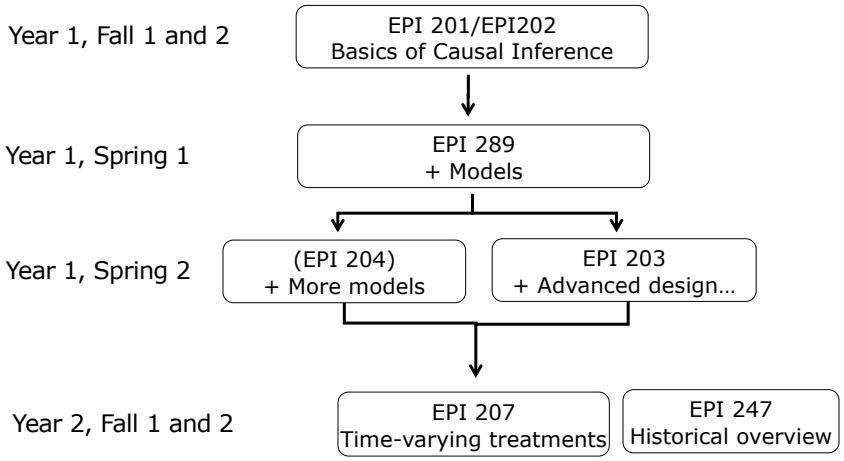
Data scientists ask three types of questions

Hernan et al. *Chance* 2019; 32(1):42-49

1. What is the incidence of heart disease in this population?
   - **Description**
2. Which individuals are at the highest risk of heart disease in this population?
   - **Prediction**
3. What would be the risk of heart disease in this population if we implement some intervention?
   - **Causal inference** (more generally, counterfactual prediction)

- Each task requires different data, methods, and subject-matter knowledge

Methods for causal inference: a key component of core epidemiology courses



EPI 201/202 Epidemiologic Methods I and II These courses set the stage

1. Dichotomous treatments: 2 levels only
   - Same in EPI289
     - "treatment and "exposure" mean the same in EPI289
   - No need to worry about dose-response curve
   - For handling of non-binary treatments, see EPI204
2. Time-fixed treatments
   - Same in EPI289
   - For handling of time-varying treatments, see EPI207
3. Data analysis mostly without models
   - On the contrary, EPI289 is all about models!

Time-fixed vs. time-varying treatments

- Time-fixed or point treatments
  - Treatment/intervention at single point in time
  - Not common in epidemiology
  - Surgery, one-dose vaccine, traffic accident, …
- Time-varying treatments
  - Treatment/intervention at multiple points in time
  - Common in epidemiology
  - Drugs, diet, exercise, screening …

EPI289: Causal inference for time-fixed dichotomous treatments **with** models

- Methods covered
  - Stratification/Regression
  - Standardization
  - Inverse probability (IP) weighting
  - G-estimation / Instrumental variables
  - Matching
- Taught via linear and logistic models
  - useful to introduce concepts and frequently used in practice
  - EPI289 does not describe the estimation procedures, e.g., maximum likelihood
- Applied to real data

EPI289: Focus on follow-up studies

- The follow-up study is the central design for causal inference
  - Randomized experiment
  - Observational cohort studies
- Other designs can be viewed as alternative ways to select persons or person-time
  - Case-control, case-base, case-cohort, case-crossover…
- Similar concepts apply to all designs

This course covers

1. Why are models necessary for causal inference?
2. Estimation of causal effects using various modeling approaches
3. Relative advantages and disadvantages of each modeling approach
4. Conditions required by each approach

This course does not cover

- ❖ whether the conditions required for causal inference are met in a particular case

- ☐ That is covered in subject-matter courses
  - ■ Cardiovascular epidemiology
  - ■ Social epidemiology
  - ■ etc.
- ☐ Expert knowledge needed for causal inference

EPI289: Outline

- ☐ Introduction to modeling
- ☐ Stratification
  - ■ Outcome regression (linear, logistic) + Propensity scores
- ☐ Standardization
  - ■ Parametric g-formula
- ☐ IP weighting
  - ■ Marginal structural models
- ☐ Instrumental variable estimation
  - ■ 2-stage least squares
- ☐ G-estimation
  - ■ Structural nested models
- ☐ Survival analysis

EPI289 designed as a complement to biostatistics courses

- EPI289 assumes students have a working knowledge of basic statistical concepts
  - e.g., variance, P-value, 95% confidence interval
- EPI289 does not describe the statistical methodology to obtain parameter estimates in linear and logistic models
  - e.g., ordinary least squares, maximum likelihood

EPI289 designed as a complement to biostatistics courses

- EPI289 focuses on conceptual issues regarding causal inference with models
  - e.g., conditions required to endow model estimates with a causal interpretation
- Historically, this has not been the emphasis of biostatistics courses

➢ To explain what I mean, let me take a detour here

Statistics and causal inference from observational data?

- Official response in the 20th century: **NO WAY!**
  - Statistics unable to aid in causal inference from observational data
- A radical disconnect
  - Mainstream statisticians avoided causal inference from observational data
  - Health and social scientists routinely used statistical methods to justify causal inferences

A refusal to tackle causal questions explicitly leads to malpractice (Hernán. *Am J Pub Health* 2018)

AJPH PUBLIC HEALTH OF CONSEQUENCE

The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data

Miguel A. Hernán, MD, DrPH

Causal inference is a core task of science. However, authors and editors often refrain from explicitly acknowledging the causal goal of research pro-

See also Galea and Vaughan, p. 602; Begg and March, p. 620; Ahern, p. 621; Chiolero, p. 622; Glymour and Hamad, p. 623; Jones and Schooling, p. 624; and Hernán, p. 625.

A familiar message from journal editors

"Dear author:

Your observational study cannot prove causation.

Please replace all references to causal effects with references to association"

- Most authors comply
- Further, most authors learn to avoid the term "causal"
  - No to causal effect, impact, benefit…
  - Yes to association, correlation, pattern, link…

Treating "causal" as a forbidden word is bad for science

- Without the term "causal"
  - Scientific goals cannot be directly stated
  - Scientific methods cannot be adequately criticized
- What's the justification for proscribing the term "causal"?
  - Association is not causation
  - i.e., there may be confounding

Of course association is not causation

- But that statement misses the point
- Suppose we want to know whether
  - daily drinking of a glass of red wine affects the 10-year risk of heart disease
- There are no randomized trials, so we use observational data
  - compare heart disease risk across people with different levels of red wine drinking over 10 years

Finding from our observational study

- Risk ratio of heart disease: 0.8
  - for 1 glass of red wine per day vs. no alcohol drinking
    - (disregard random variability and measurement error)
- 0.8 measures the **association** between wine intake and heart disease
  - Strictly speaking, it means "drinkers of 1 glass of wine per day have, on average, a 20% lower risk of heart disease than nondrinkers"

Why is the risk ratio 0.8?

- Not necessarily because drinking 1 glass of wine lowers the risk of heart disease by 20%
- Perhaps the kind of people who drink 1 glass of wine per day would have a lower risk of heart disease even if they didn't drink wine
  - wealthier, better access to health care…
- 0.8 may be a valid estimate of association, but a confounded estimate of causal effect

"Association does not imply causation in observational studies"

- Not a scientific statement but a logical one
- The statement "Your causal estimate may be seriously confounded" cannot be proven wrong
  - No matter how much observational data we collect
- But avoiding causal language doesn't solve this problem
  - It makes it worse

Risk ratio of heart disease is 0.8 for 1 glass of wine per day vs. no drinking

- If we were truly interested in the association, no need to adjust for anything
  - No confounding for associations
- If we are interested in the causal effect, need to adjust for confounders
  - variables that predict both wine drinking and heart disease
  - Identified and selected using expert knowledge

But there is no guarantee that all confounders will be identified!

- Therefore the causal effect estimated via an adjusted association may be confounded
  - Association is not causation
- We have come full circle
  - There is no guarantee the associational estimate can be causally interpreted, but an informed scientific discussion requires that we first acknowledge the causal goal of the data analysis

Conflating the means and the ends

- The **goal** of our study was to quantify the causal effect of wine drinking on heart disease
  - Not the association between them
- We attempt to achieve that goal by computing associations
  - If truly randomized trial, we feel more confident
  - If observational analysis emulating a target trial, we feel less confident
- Computing associations is just a method for causal inference, not the goal itself

Without causally explicit language, means and ends get hopelessly conflated

- The result is inconsistency in a scientific manuscript
  - Authors will repeatedly assure you that they are just computing associations for much of the paper
    - The association between wine and heart disease is 0.8
    - The risk of heart disease is 20% lower in wine drinkers
  - And then, without warning, they will make causal claims
    - Wine drinking may lower the risk of heart disease
    - We recommend moderate wine drinking
- Why not accept the causal goal from the start?

OPEN ACCESS

Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study

Carlos A Celis-Morales,[1] Donald M Lyall,[2] Paul Welsh,[1] Jana Anderson,[2] Lewis Steell,[1] Yibing Guo,[1] Reno Maldonado,[1] Daniel F Mackay,[2] Jill P Pell,[2] Naveed Sattar,[1] Jason M R Gill[1]

[1]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow G12 8TA, UK
[2]Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK
Correspondence to: J M R Gill
jason.gill@glasgow.ac.uk
Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2017;357:j1456
http://dx.doi.org/10.1136/bmj.j1456

Accepted: 16 March 2017

ABSTRACT

OBJECTIVE
To investigate the association between active commuting and incident cardiovascular disease (CVD), cancer, and all cause mortality.

DESIGN
Prospective population based study.

SETTING
UK Biobank.

PARTICIPANTS
263 540 participants (106 674 (52%) women; mean age 52.6), recruited from 22 sites across the UK. The exposure variable was the mode of transport used (walking, cycling, mixed mode *v* non-active (car or public transport)) to commute to and from work on a typical day.

MAIN OUTCOME MEASURES
Incident (fatal and non-fatal) CVD and cancer, and deaths from CVD, cancer, or any causes.

RESULTS
2430 participants died (496 were related to CVD and 1126 to cancer) over a median of 5.0 years (interquartile range 4.3-5.5) follow-up. There were 3748 cancer and 1110 CVD events. In maximally adjusted models, commuting by cycle and by mixed mode cause mortality (cycling hazard ratio 0.59, 95% confidence interval 0.42 to 0.83, P=0.002; mixed mode cycling 0.76, 0.58 to 1.00, P<0.05), cancer incidence (cycling 0.55, 0.44 to 0.69, P<0.001; mixed mode cycling 0.64, 0.45 to 0.91, P=0.01), and cancer mortality (cycling 0.60, 0.40 to 0.90, P=0.01; mixed mode cycling 0.68, 0.57 to 0.81, P<0.001). Commuting by cycling and walking were associated with a lower risk of CVD incidence (cycling 0.54, 0.33 to 0.88, P=0.01; walking 0.73, 0.54 to 0.99, P=0.04) and CVD mortality (cycling 0.48, 0.25 to 0.92, P=0.03; walking 0.64, 0.45 to 0.91, P=0.01). No statistically significant associations were observed for walking commuting and all cause mortality or cancer outcomes. Mixed mode commuting including walking was not noticeably associated with any of the measured outcomes.

CONCLUSIONS
Cycle commuting was associated with a lower risk of CVD, cancer, and all cause mortality. Walking commuting was associated with a lower risk of CVD independent of major measured confounding factors. Initiatives to encourage and support active commuting could reduce risk of death and the burden of important chronic conditions.

Introduction

27

OPEN ACCESS

Association bet
cardiovascular d
cohort study

Carlos A Celis-Morales,[1]
Reno Maldonado,[1] Dani

[1]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow G12 8TA, UK
[2]Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK
Correspondence to: J M R Gill
jason.gill@glasgow.ac.uk
Additional material is published online only. To view please visit the journal online.
Cite this as: *BMJ* 2017;357:j1456
http://dx.doi.org/10.1136/bmj.j1456
Accepted: 16 March 2017

ABSTRACT
OBJECTIVE
To investigate the associati
commuting and incident ca
cancer, and all cause morta
DESIGN
Prospective population bas
SETTING
UK Biobank.
PARTICIPANTS
263 540 participants (106 6
52.6), recruited from 22 site
exposure variable was the m
(walking, cycling, mixed mo
public transport)) to commu
typical day.
MAIN OUTCOME MEASURES
Incident (fatal and non-fata
deaths from CVD, cancer, o
RESULTS
2430 participants died (49
1126 to cancer) over a medi
(interquartile range 4.3-5.5)
cancer and 1110 CVD events. In maximally adjusted
models, commuting by cycle and by mixed mode

Introduction

thebmj

BMJ 2017;357:j1944 doi: 10.1136/bmj.j1944 (Published 19 April 2017)

RESEARCH NEWS

Check for updates

Cycling to work has substantial health benefits, study finds

Ingrid Torjesen

London

Cycling to work is linked to a substantial decrease in the risk of developing and dying from cancer or heart disease, a study published in *The BMJ* has found.[1]

Walking was also associated with a lower risk of cardiovascular disease (CVD), but the risk of death from cancer was no lower

The researchers concluded that "the findings, if causal, suggest population health may be improved by policies that increase active commuting, particularly cycling, such as the creation of cycle lanes, cycle hire or purchase schemes, and better provision for cycles on public transport."

28

OPEN ACCESS

Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study

Carlos A Celis-Morales,[1] D
Reno Maldonado,[1] Danie

[1]Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow G12 8TA, UK
[2]Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK
Correspondence to: J M R Gill
jason.gill@glasgow.ac.uk
Additional material is published online only. To view please visit the journal online.
Cite this as: *BMJ* 2017;357:j1456
http://dx.doi.org/10.1136/bmj.j1456
Accepted: 16 March 2017

ABSTRACT
OBJECTIVE
To investigate the associatio
commuting and incident card
cancer, and all cause mortali
DESIGN
Prospective population base
SETTING
UK Biobank.
PARTICIPANTS
263 540 participants (106 67
52.6), recruited from 22 sites
exposure variable was the m
(walking, cycling, mixed mod
public transport)) to commut
typical day.
MAIN OUTCOME MEASURES
Incident (fatal ar
deaths from CVD
RESULTS
2430 participant
1126 to cancer) o
(interquartile ra
cancer and 1110
models, commut

thebmj

BMJ 2017;357:j1740 doi: 10.1136/bmj.j1740 (Published 2017 April 20) Page 1 of 2

EDITORIALS

Check for updates

Active commuting is beneficial for health

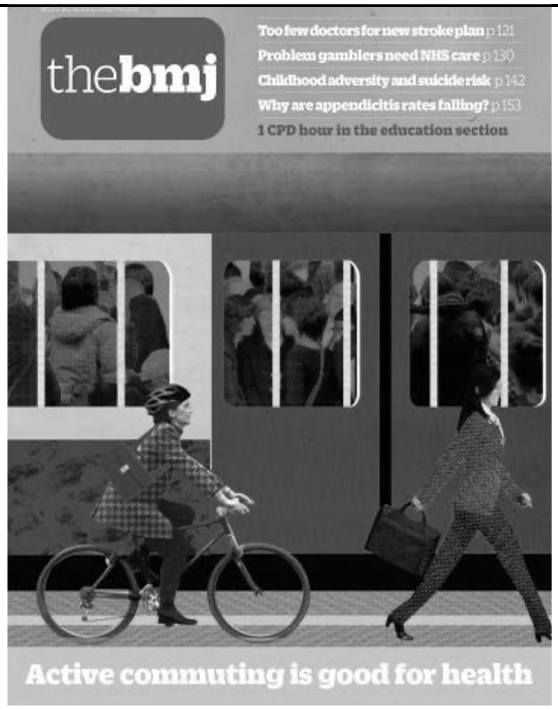
Governments should do all they can to encourage commuters to cycle or walk

Lars Bo Andersen *professor*

Department of Teacher Education and Sport, Western Norwegian University of Applied Sciences, Bergen, Norway

"The findings from this study are a clear call for political action on active commuting, which has the potential to improve public health by preventing common (and costly) non-communicable diseases"

29



The word "causal" everywhere except in the actual article.

Is this defensible?

Is this scientific?

30

In scientific papers, the term "causal effect" is appropriate in

- Title, Introduction, Methods
  - Describe the causal effect of interest by specifying the target trial
  - Describe the proposed emulation procedure
- Discussion
  - Provide arguments for and against the causal interpretation of the findings
- The only section of the paper in which "causal effect" has no place is the Results section
  - Present findings without interpreting them

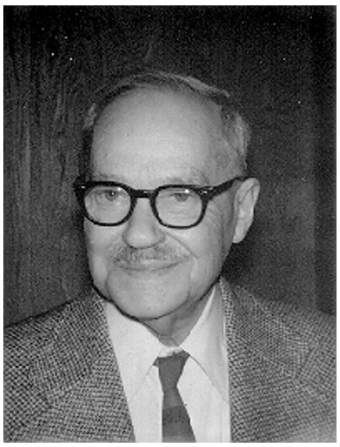
Practical implications of embracing the word "causal"

(besides enhancing scientific communication and transparency)

1. Better causal questions
   - Specify the target trial that would answer the causal question of interest
2. Better causal methods
   - Identify and adjust for important confounders

Fortunately, some statisticians challenged the official view of statistics regarding causal inference

- Neyman (1923)
  - Effects of point or fixed treatments in randomized experiments
- Rubin (1974)
  - Effects of point or fixed treatments in randomized and observational studies
- Robins (1986)
  - Effects of time-varying treatments in randomized and observational studies



Back to EPI289 Organization of the course

- Lectures: Mon, Wed
  - 9:45am-11:15am EST
- Labs: Wed
  - 11:30am-1:00pm **or** 2:00-3:30pm **or** 3:45-5:15pm EST
  - No lab last week
- One optional seminar
- Office hours (optional)
  - 6 time slots per week
  - See course site for locations, times

Lectures

- Feel free (and encouraged!) to ask questions
- We will have frequent real-time polls
- Your responses will NOT be used for grading purposes

Labs (TF-led)

- Homework review, discussion
- Weekly homeworks will revolve around the analysis of NHANES data
  - Each week you will be asked to estimate the same causal effect using a different method
- Think of homeworks as a learning experience
  - Sometimes not necessarily right or wrong answers, but well- or ill-reasoned answers

Assignments (I)
Weekly homeworks

- Homeworks due via Canvas by the start of class on Wednesdays
  - Late homeworks will be penalized
  - If you are still waitlisted, turn in your homeworks just in case
- You are encouraged to **work in groups** to discuss the homeworks, but you must turn in individual answers
- You are expected to create your own answer sheets

Assignment (II)
Take-home final exam

- Will revolve around the analysis of a real data set
- Important
  - Due on last Wednesday at start of class
  - **Work individually**

Course materials

- Materials from EPI 201/202
  - Including videos from HarvardX Causal Diagrams course
- Class notes
  - Posted to course site before class
- Selected papers
  - Required and recommended
  - Posted to course site
- "*Causal Inference: What If*" book. Part II

R: official computing language of EPI289

- We will provide R code
  - but we will not teach R
  - SAS code is also available for students on the course site, but we will not review SAS code during class, labs or office hours
- If learning R, resources on web site:
  - R reference document
- Computer-based assignments for first lab
  - Yes, the day after tomorrow

GAI policy

- Permitted to use GAI tools (e.g., ChatGPT) to debug R code for statistical analysis
  - Must be appropriately acknowledged and cited
  - Your responsibility to assess the applicability and accuracy of code
- Not permitted to use GAI tools to generate written text for course assessments
  - Including but not limited to interpretations, assumptions, and explanations
- Please see course syllabus for full policy on use of GAI tools in EPI289

What to do when questions arise in EPI289

1. Ask questions to Barbra or Joy during the lectures
   - Do not leave the room with unanswered questions
2. Ask questions to your TF during the lab sessions
   - Do not leave the room with unanswered questions
3. Ask questions to any TF during office hours
   - You can try up to 6 office hours if necessary
   - See course Canvas or syllabus for times and location
4. Post questions to the discussion board