

Pre-conference Workshop: Introduction to Causal Inference for Epidemiologists (I). XXXVII-SEE, Oviedo, Spain 2019

Miguel Angel Luque Fernandez

<https://maluque.netlify.com/>
<https://ccci.netlify.com/>

PART I: IDENTIFICATION and IGNORABILITY

Justification, Potential Outcomes, Ignorability, and DAGs

Public Health as Scientific Discipline: subdisciplines

The screenshot shows a web browser window for the Harvard T.H. Chan School of Public Health. The URL is https://www.hsph.harvard.edu/departments/. The page features a navigation bar with links for Prospective Students, Current Students, Alumni, Faculty & Staff, and Friends & Supporters. Below this is a logo for HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH. The main content area is titled "Academic Departments, Divisions and Centers" and includes a sidebar with links to various academic programs and administrative offices. A search bar is located at the top right of the main content area.

Departments | Harvard T.H.C × +

https://www.hsph.harvard.edu/departments/ 50% Search

Most Visited 38th Annual Conferen... Home - Research Parti... Surveillance Research ... Tutoriales - Big Data y... A Primer in Econon

INFORMATION FOR: Prospective Students Current Students Alumni Faculty & Staff Friends & Supporters

HARVARD
T.H. CHAN SCHOOL OF PUBLIC HEALTH

ABOUT | FACULTY & RESEARCH | ADMISSIONS & AID | ACADEMICS | EXECUTIVE/CONTINUING ED | NEWS

Departments

> Departments

DEPARTMENTS

Search this section

Academic Departments, Divisions and Centers

Academic Departments

- ▼ Biostatistics
- ▼ Environmental Health
- ▼ Epidemiology
- ▼ Genetics and Complex Diseases
- ▼ Global Health and Population
- ▼ Health Policy and Management
- ▼ Immunology and Infectious Diseases
- ▼ Nutrition
- ▼ Social and Behavioral Sciences

Epidemiology as scientific method: areas of concentration

S Department of Epidemiology x +

Most Visited 38th Annual Conference Home - Research Part... Surveillance Research ... Tutoriales - Big Data y... A Primer in Economet... <https://www.hsph.harvard.edu/epidemiology/> 67% Search

INFORMATION FOR: Prospective Students Current Students Alumni Faculty & Staff Friends & Supporters

DEPARTMENT OF EPIDEMIOLOGY

Search this section 

Home

Department, Staff & Faculty

Areas of Epidemiology

- Cancer Epidemiology
- Cardiovascular Epidemiology
- Clinical Epidemiology
- Environmental and Occupational Epidemiology
- Epidemiologic Methods
- Epidemiology of Aging
- Infectious Disease Epidemiology
- Genetic Epidemiology and Statistical Genetics
- Neuro-Psychiatric Epidemiology
- Nutritional Epidemiology
- Pharmacoepidemiology
- Reproductive, Perinatal, and Pediatric Epidemiology

Welcome To the World Renowned Epidemiology Department

 Albert Hofman MD, Ph.D
Stephen B. Kay Family Professor of Public Health and Clinical Epidemiology
Chair, Department of Epidemiology,
Harvard T.H. Chan School of Public Health

Additional Information

Welcome to the Department of Epidemiology at the Harvard T.H. Chan School of Public Health. We study the frequency and determinants of disease in humans, a fundamental science of public health. In addition to pursuing groundbreaking research initiatives, we educate and prepare future medical leaders and practitioners as part of our mission to ignite changes in the quality of health across the world.

Innovative Educational Programs: Onsite and Online

Data Science Initiative

co-directors of newly launc... +

https://news.harvard.edu/gazette/story/2017/03/co-directors-of-newly-launched-data-science-initiative-harvard-university/ 80% Search

Visited 38th Annual Conferenc... Home - Research Part... Surveillance Research ... Tutoriales - Big Data y... A Primer in Econometr... GESTIÓN DE PROYECT... DeepL Translate

g HOME Search harvard.edu Photographic Services Resources for Journalists HPAC

CAMPUS & COMMUNITY

- Awards
- Commencement
- Faculty
- Harvard News
- Harvard Traditions
- In the Community
- News by School
- Obituaries
- On Campus

Staff & Administration

- Staff News

ARTS & CULTURE

SCIENCE & HEALTH

NATIONAL & WORLD AFFAIRS

ATHLETICS

HARVARD EVENTS

GAZETTE TOPICS

Subscribe to the Daily Gazette

■ CAMPUS & COMMUNITY > STAFF & ADMINISTRATION

Data science for a new era

A Q&A with co-directors of emerging Data Science Initiative

March 28, 2017 | ✓



Kris Snibbe/Harvard Staff Photographer

Harvard University just announced the launch of its [Data Science Initiative](#), a program to harness the vast expertise and innovations that are occurring in disciplines as diverse as medicine, law, policy, and computer science.

Initiative co-directors [Francesca Dominici](#), professor of biostatistics at the [Harvard T.H. Chan School of Public Health](#), and [David C. Parkes](#), George F. Colony Professor and area dean for computer science at the [Harvard John A. Paulson School of Engineering and Applied Sciences](#), are enthusiastic about the work ahead.

In a Q&A session, Dominici and Parkes talked with The Gazette about their work ahead.

POPULAR

Data Science Programmes

A screenshot of a web browser window showing the Harvard Data Science website. The address bar shows 'data-science.harvard.edu/education'. The page title is 'Data Science Programmes'. The navigation menu includes links for 'HOME', 'ABOUT', 'NEWS', 'COMMUNITY', 'EDUCATION' (which is highlighted), 'EVENTS', and 'APPLY'. There are also links for 'For Graduate Students' and 'For Undergraduates'. The main content area discusses data science education for graduate students.

At Harvard University, data science education for graduate students is a rapidly growing field. A growing community of data scientists, research scientists, and methodologists is developing in several key academic areas on campus: Engineering/Statistics, Medicine, and Public Health.

Master's Degree Programs in Data Science

Institute for Applied Computational Science | John A. Paulson Harvard School of Engineering and Applied Sciences

[Master of Science in computational science and engineering](#)

[Master of Engineering in computational science and engineering](#)

Biomedical Informatics | Harvard Medical School

[Master of Biomedical Informatics Program](#)

Health Data Science | Harvard T.H. Chan School of Public Health

[Master's degree program](#)

Data Science | Faculty of Arts and Sciences

[Master of Science in Data Science launching in 2018](#)

PhD Programs in Data Science

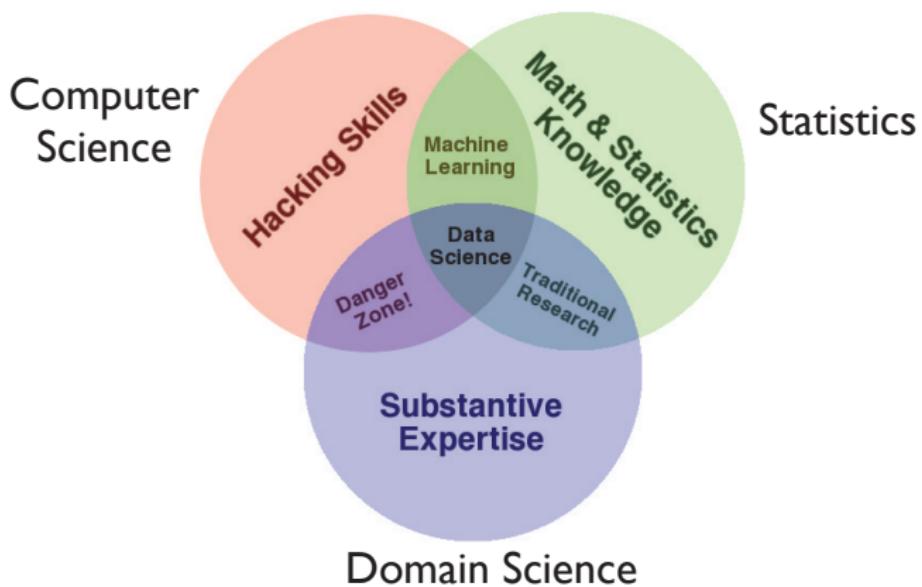
Bioinformatics | Harvard Medical School

[PhD in Bioinformatics and Integrative Genomics \(BIG\)](#)

Doctoral Degree Secondary Field Opportunities



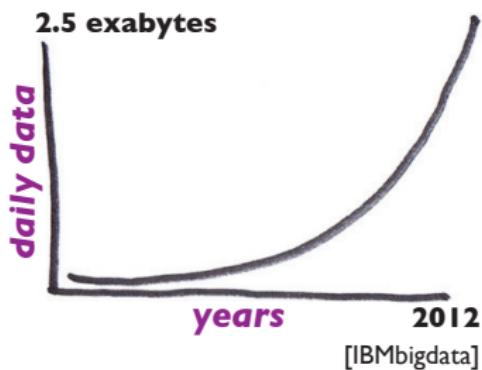
Data Science



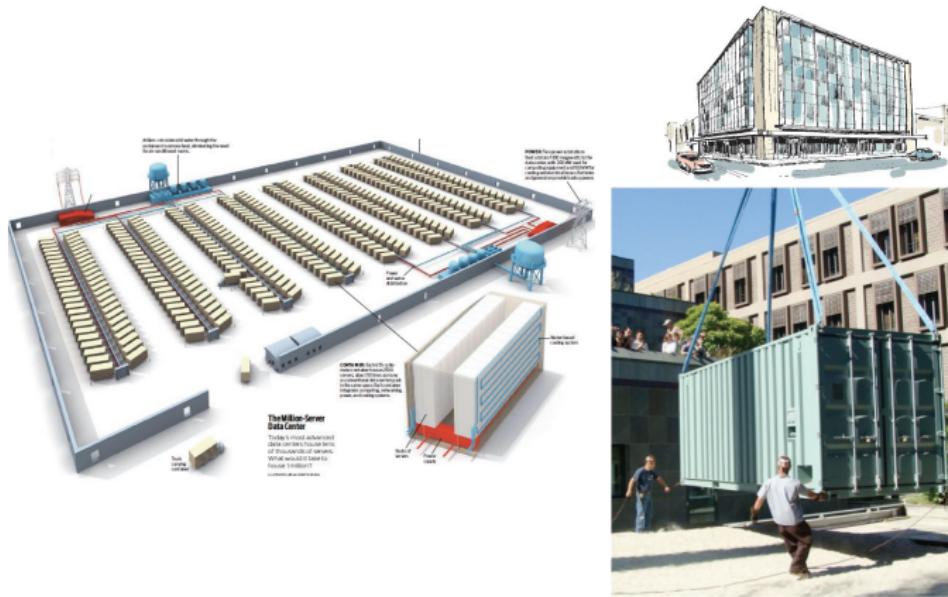
Drew Conway

Data Science and Big Data: Volume

Big Data



Commodity Computing



Michael Franklin, UC Berkeley

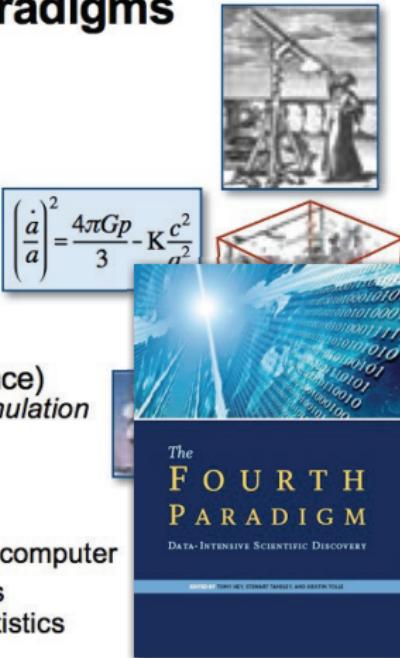
Smarter Devices



Michael Franklin, UC Berkeley

Science Paradigms

- Thousand years ago:
science was empirical
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a computational branch
simulating complex phenomena
- Today: **data exploration (eScience)**
unify theory, experiment, and simulation
 - Data captured by instruments
or generated by simulator
 - Processed by software
 - Information/knowledge stored in computer
 - Scientist analyzes database/files
using data management and statistics



Jim Gray, Microsoft

Data Science the sexiest job

“By 2018, the US could face a shortage of up to 190,000 workers with analytical skills”

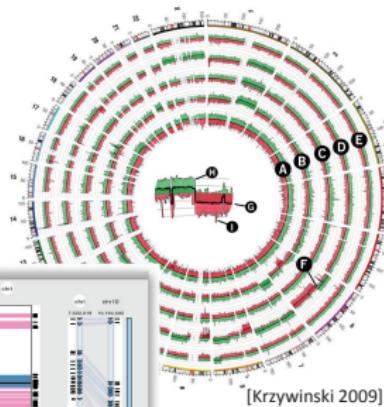
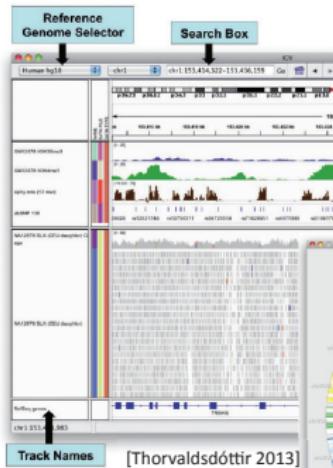
McKinsey Global Institute

“The sexy job in the next 10 years will
~~be statisticians.~~” *Data Scientists?* Epidemiologists?

Hal Varian, Prof. Emeritus UC Berkeley
Chief Economist, Google

So, how about Epidemiology?

Genome Visualization



[Meyer 2009]

So, how about Epidemiology?

electronic charting

health record

digital

patient

nursing

doctor



So, how about Epidemiology?



Journal of Clinical Epidemiology 58 (2005) 323–337

REVIEW ARTICLE

A review of uses of health care utilization databases for epidemiologic research on therapeutics

Sebastian Schneeweiss*, Jerry Avorn

Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street (suite 3030), Boston, MA 02120, USA

Accepted 16 October 2004

Abstract

Objective: Large health care utilization databases are frequently used in variety of settings to study the use and outcomes of therapeutics. Their size allows the study of infrequent events, their representativeness of routine clinical care makes it possible to study real-world effectiveness and utilization patterns, and their availability at relatively low cost without long delays makes them accessible to many researchers. However, concerns about database studies include data validity, lack of detailed clinical information, and a limited ability to control confounding.

Study Design and Setting: We consider the strengths, limitations, and appropriate applications of health care utilization databases in epidemiology and health services research, with particular reference to the study of medications.

Conclusion: Progress has been made on many methodologic issues related to the use of health care utilization databases in recent years, but important areas persist and merit scrutiny. © 2005 Elsevier Inc. All rights reserved.

Keywords: Utilization databases; Claims data; Therapeutics; Pharmaco-epidemiology; Confounding (epidemiology); Adverse drug reactions; Drug utilization

1. Introduction

It is widely accepted that randomized clinical trials (RCT) cannot provide all necessary information about the safe and effective use of medicines at the time they are marketed. This stems from the inherent limitations of RCTs during drug development: They usually have a small sample size that often under-represents vulnerable patient groups, and they focus on short-term efficacy and safety in a controlled environment that is often far from routine clinical practice. Moreover, the RCT outcome sufficient to win marketing approval—short-term improvement in a surrogate marker compared with the effect of placebo—often fails to answer the

and put them into context of the natural history of the condition they are designed to treat [4].

Although pharmacoepidemiology makes use of all epidemiologic study designs and data sources, in recent years there has been enormous growth in the use of large health care databases [5]. These are made up of the automated electronic recording of filled prescriptions, professional services, and hospitalizations; such data are increasingly collected routinely for the payment and administration of health services. Beyond this, electronic medical records often contain detailed clinical information, patients' reports of symptoms, the findings of physical examinations, and the results

Conclusion

"(...) Increasing availability in electronic medical records of even more detailed clinical information, such as the medical history and the results of diagnostic tests, will further enhance the validity and versatility of the use of **electronic health records** (...)."

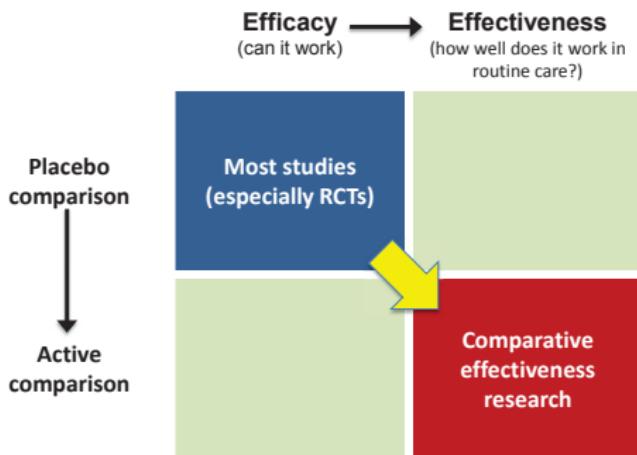
CER, defined

...is the generation and synthesis of evidence that
**compares the benefits and harms of alternative
methods to prevent, diagnose, treat, and
monitor a clinical condition or to improve the
delivery of care.**

Source: Institute of Medicine, *Initial National Priorities for Comparative Effectiveness Research*, 2009.

CER is different

How CER is different



SOURCE: Academy Health. "A first look at the volume and cost of comparative effectiveness research in the United States." Academy Health, 2009. http://www.old.academyhealth.org/files/FileDownloads/AH_Monograph_09FINAL7.pdf

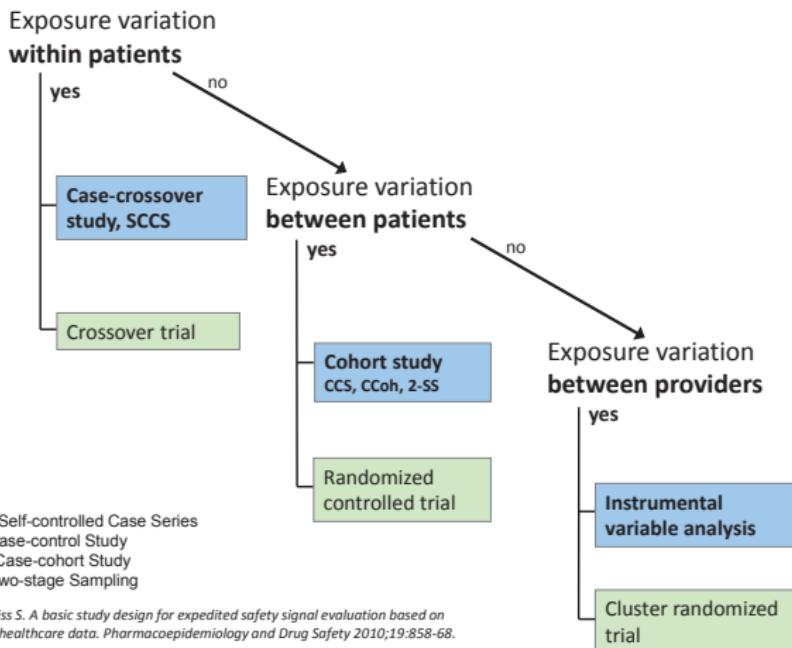
What CER seeks to do

	TYPICAL RCTs	NEEDS OF DECISION MAKERS
Comparator	Placebo or usual care	Active
Patient population	Highly selected	Representative of typical practice
Outcome measures	Surrogate	Patient centered
Follow-up time	Short	Long
Cost	High	Moderate
Speed	Slow	Faster

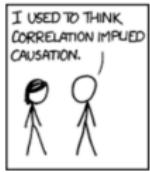
¹ Harvard Catalyst Comparative Effectiveness Research Course
<https://catalyst.harvard.edu/services/cer/>

CER is about New Epidemiological Methods

Design choice: Source of exposure variation



CER is about Causal Inference

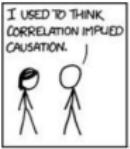


CAUSAL INFERENCE IN STATISTICS

A Primer

Judea Pearl
Malcolm Glymour
Nicholas P. Jewell

WILEY



A Research Agenda on Causal Inference

Problems

• How to distinguish causal from non-causal associations in social sciences (e.g., predictions and causal relations)

• Existing ML approaches to causal inference are not directly applicable to the complex causal processes in social sciences

• Inference rules are not yet formalized

Predictor

• Cred

• High birth weight

• Newborn

• Days in hospital

• Male

• First term

• Age

• Mother

• Father

• White

• Married

• Born at birth

• Less than high school

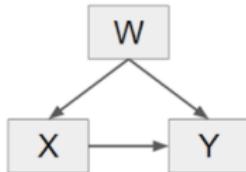
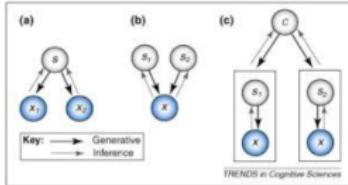
• High school graduate

• College graduate

• Worked during pregnancy

• Maternal age at birth

• Age at death



Causal Inference

- Causal inference is essentially about control and explanation.
- Good control should require good predictive models anyway.
- Explanation is not about the future, but counterfactual events in the past.
- How to solve these problems?

Headlines

- Levels of causality
- Definitions
- Koch's postulates (1877)
- Hill's criteria (1965)
- Susser's criteria (1988, 1991)

Causal Inference

• A causal inference is a statement about why something happens.

• A causal inference therefore states the existence of a relationship between at least two variables.

• The dependent variable measures that variation which we would like to explain (find a cause for).

• Also called **X** or the "dependent" or "responsible" variable.

• The independent variable measures that variation which we think explains variation in the dependent variable.

• Also called **K** or the "treatment" or "study" variable.

Estimating causal effects from observational data

Why is causal inference hard?

• Simpson's paradox

The language of graphical models

• Backdoor criterion

• Frontdoor criterion

Common approaches for causal inference

• Covariate adjustment

• Mechanism-based

• Natural Experiments

Example: Estimating causal impact of recommender systems

Model for Causal Inference

For causal questions, we want to know:

• What would happen if a policy-maker changes something?

• Potential outcomes notation

• If you change this, what would happen?

• For binary treatment, there are two cases

• Administer a drug, change miles driven

• Function of interest: mapping from causes to effects

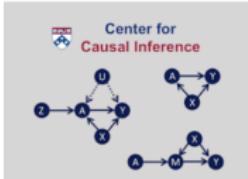
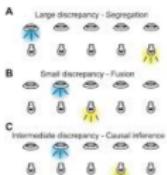
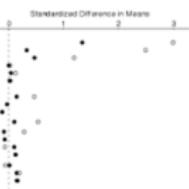
• We do not see the same outcome

Units of study typically have:

• These would not change with time

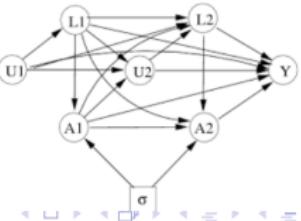
• E.g., we don't compromise

• We change minimum wage per hour



Causal Inference Diagrams

Parametric Approaches



C

M

O

E

CER is about Causal Inference

causal inference - Google Search Deconstructing the smoking Deconstructing the smoking collapsibility odds ratio - Google Search Targeted Maximum Likelihood Estimation

https://migariane.github.io/TMLE.nb.html

Most Visited 38th Annual Conference Home - Research Part... Surveillance Research ... Tutoriales - Big Data y... A Primer in Economet... GESTIÓN DE PROYECTOS DeepL Translator LOGSE

1 Introduction
2 The G-Formula and ATE estimation
3 TMLE
4 Structural causal framework
4.1 Direct Acyclic Graph (DAG)
4.2 DAG interpretation
5 Causal assumptions
5.1 CMI or Randomization
5.2 Positivity
5.3 Consistency or SUTVA
6 TMLE flow chart
7 Data generation
7.1 Simulation
7.2 Data visualization
8 TMLE simple implementation
8.1 Step 1: $Q_0(A, W)$
8.2 Step 2: $g_0(A, W)$
8.3 Step 3: HAW and ϵ
8.4 Step 4 \bar{Q}_0^* from \bar{Q}_0^0 to \bar{Q}_0^1

Targeted Maximum Likelihood Estimation for a Binary Outcome: Tutorial and Guided Implementation

By: Miguel Angel Luque Fernandez

June 20th, 2017

Code

Migariane



1 Introduction

During the last 30 years, **modern epidemiology** has been able to identify significant limitations of classic epidemiologic methods when the focus is to explain the main effect of a risk factor on a disease or outcome.

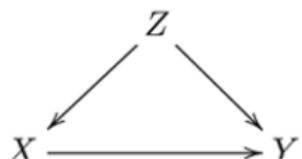
Causal Inference based on the **Neyman-Rubin Potential Outcomes Framework** (Rubin, 2011), first introduced in Social Science by Donal Rubin (Rubin, 1974) and later in Epidemiology and Biostatistics by James Robins (Greenland and Robins, 1986), has provided the theory and statistical methods needed to overcome recurrent problems in observational epidemiologic research, such as:

1. non-collapsibility of the odds and hazard ratios,
2. impact of paradoxical effects due to conditioning on colliders,
3. selection bias related to the vague understanding of the effect of time on exposure and outcome and,
4. effect of time-dependent confounding and mediators,
5. etc.

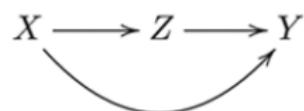
Causal effects are often formulated regarding comparisons of potential outcomes, as formalised by Rubin (Rubin, 2011). Let A denote a binary exposure, W a vector of potential confounders, and Y a binary outcome. Given A , each individual has a pair of potential outcomes: the outcome when exposed, denoted Y_1 , and the outcome when unexposed, Y_0 . These quantities are referred to as **potential outcomes** since they are hypothetical, given that it is only possible to observe a single realisation of the outcome for an individual; we observe Y_1 only for those in the exposure group and Y_0 only for those in the unexposed group (Rubin, 1974). A common causal estimand is the **Average Treatment Effect (ATE)**, defined as $E[Y_1 - Y_0]$.

CER is about Causal Inference

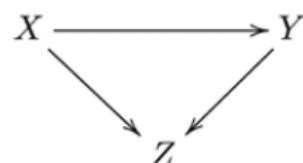
A



B



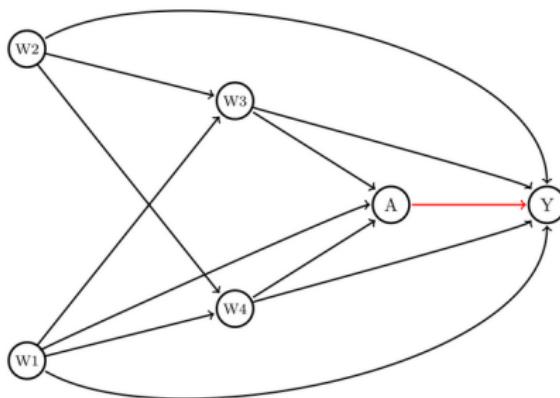
C



CER is about Causal Inference

Direct Acyclic Graph (DAG)

Under conditional exchangeability: $Y(0), Y(1) \perp\!\!\!\perp A|W$
ATE = $E[Y|A = 1; W] - E[Y|A = 0; W]$



Y = Mortality; A = Chemotherapy vs. Chemotherapy & Radiotherapy; W_1 = Sex; W_2 = Age; W_3 = TNM-Stage; W_4 = Comorbidities

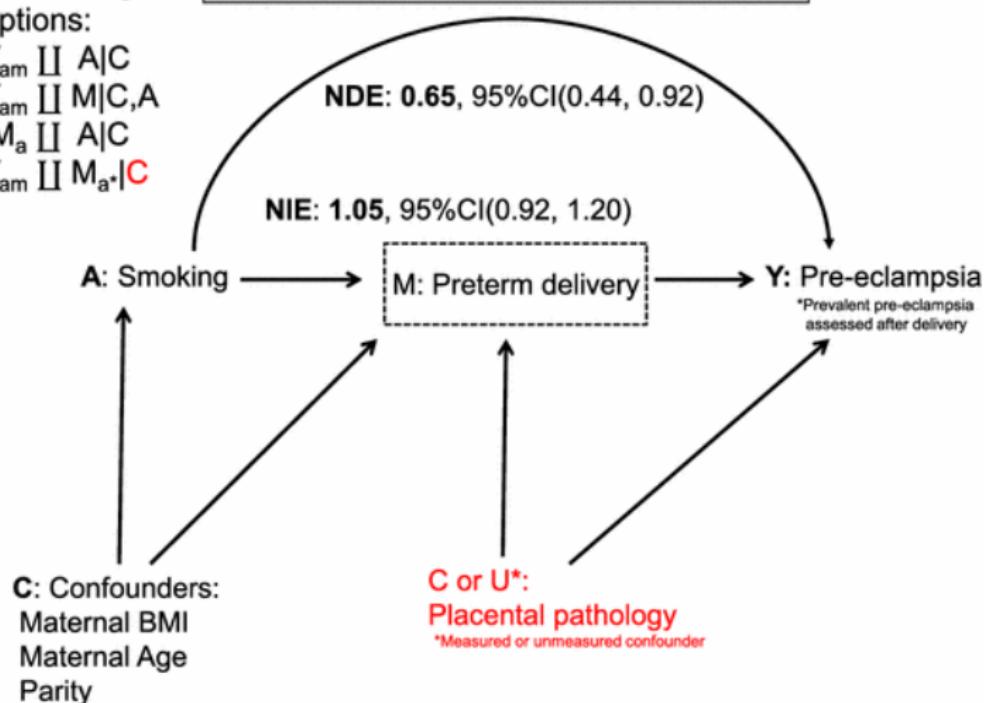
² Data-Adaptive Estimation for Double-Robust Methods in Population-Based Cancer Epidemiology: Risk differences for lung cancer mortality by emergency presentation (2017). AJE.
<https://academic.oup.com/aje/article/doi/10.1093/aje/kwx317/4110407>

CER is about Causal Inference

Mediation analysis: Marginal Total Effect: **0.68, 95%CI(0.45, 0.97)**

Assumptions:

- (1) is $Y_{am} \perp\!\!\!\perp A|C$
- (2) is $Y_{am} \perp\!\!\!\perp M|C,A$
- (3) is $M_a \perp\!\!\!\perp A|C$
- (4) is $Y_{am} \perp\!\!\!\perp M_a|C$



³ Luque-Fernandez, M.A., Zoega, H., Valdimarsdottir, U. et al. Eur J Epidemiol (2016) 31: 613. <https://doi.org/10.1007/s10654-016-0139-5>

CER is about Causal Inference



Arvid Sjölander, Elisabeth Dahlqvist, and Johan Zetterqvist

Abstract: It is well known that the odds ratio is noncollapsible, in the sense that conditioning on a covariate that is related to the outcome typically changes the size of the odds ratio, even if this covariate is unrelated to the exposure. The risk difference and risk ratio do not have this peculiar property; we say that the risk difference and risk ratio are collapsible. However, noncollapsibility is not unique for the odds ratio; the rate difference and rate ratio are generally noncollapsible as well. This may seem paradoxical, since the rate can be viewed as a risk per unit time, and thus one would naively suspect that the rate difference/ratio should inherit collapsibility from the risk difference/ratio. Adding to the confusion, it was recently shown that the exposure coefficient in the Aalen additive hazards model is collapsible. This may seem to contradict the fact that the rate difference is generally noncollapsible, since the exposure coefficient in the Aalen additive hazards model is a rate difference. In this article, we use graphical arguments to explain why the rate difference/ratio does not inherit collapsibility from the risk difference/ratio. We also explain when and why the exposure coefficient in the Aalen additive hazards model is collapsible.

(Epidemiology 2016;27: 356–359)

When studying the association between an exposure X and an outcome Y , it is common to adjust for additional covariates Z in the analysis. For binary variables, the conditional (on Z) odds ratio

$$\frac{\text{Pr}(Y = 1 | X = 1, Z)\text{Pr}(Y = 0 | X = 0, Z)}{\text{Pr}(Y = 0 | X = 1, Z)\text{Pr}(Y = 1 | X = 0, Z)}$$

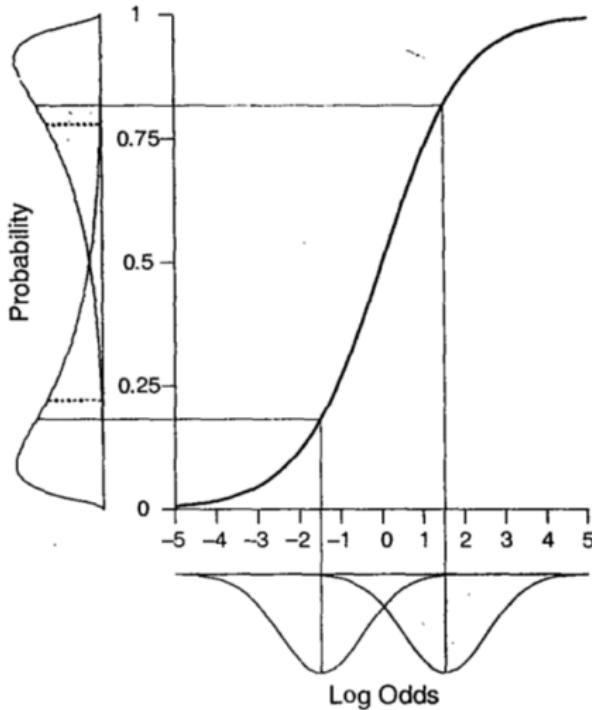
that the conditional odds ratio is constant across levels of Z (e.g., in logistic regression with main effects only).

Most epidemiologists would not be surprised to find that the conditional odds ratio is different from the unadjusted marginal (over Z) odds ratio

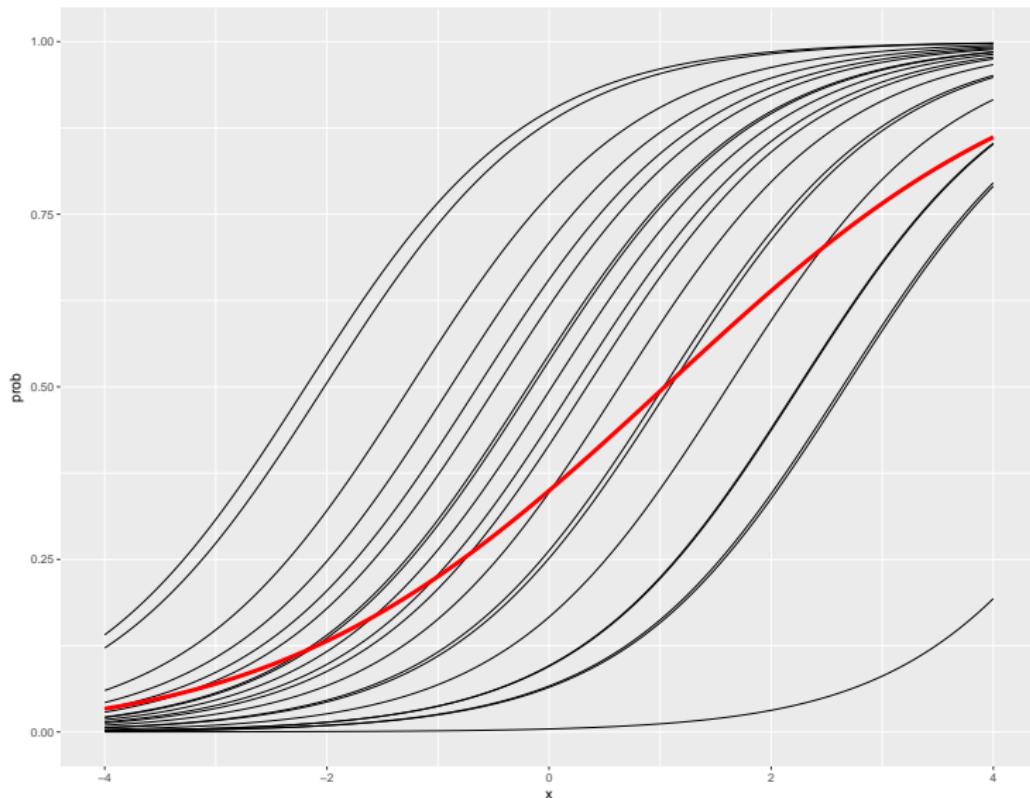
$$\frac{\text{Pr}(Y = 1 | X = 1)\text{Pr}(Y = 0 | X = 0)}{\text{Pr}(Y = 0 | X = 1)\text{Pr}(Y = 1 | X = 0)}$$

One explanation for a discrepancy between the conditional and marginal odds ratio could be that Z is a confounder (Fig. 1A); this would typically be the argument for adjusting for Z in the first place. Other explanations could be that Z is a mediator (Fig. 1B) or a collider (Fig. 1C). All these explanations require that Z is associated with both X and Y . However, the conditional odds ratio may differ from the marginal odds ratio even when Z is independent of X . To see that this behavior is rather counterintuitive, suppose that we carry out a randomized trial, so that confounding is eliminated by design. Suppose that we first calculate the marginal exposure-outcome odds ratio and find that this is equal to two. Suppose that we next calculate the exposure-outcome odds ratio for men and women separately, and find that these are both equal to three. By randomization, all these odds ratios can be interpreted as causal effects. Thus, in this example, the causal effect is three for men and three for women, but only two for men and women pooled together, all effects measured on the odds ratio scale. This numerical artifact is often referred to as noncollapsibility.¹ Neuhäusl and Jewell² showed that the mar-

CER is about Causal Inference



CER is about Causal Inference



CER is about Causal Inference

The Hazards of Hazard Ratios

Miguel A. Hernán

[Author information](#) ► [Copyright and License information](#) ►

The publisher's final edited version of this article is available at [Epidemiology](#)

This article has been corrected. See the correction in volume 22 on page 134.

See other articles in PMC that [cite](#) the published article.

The hazard ratio (HR) is the main, and often the only, effect measure reported in many epidemiologic studies. For dichotomous, non-time-varying exposures, the HR is defined as the hazard in the exposed groups divided by the hazard in the unexposed groups. For all practical purposes, hazards can be thought of as incidence rates and thus the HR can be roughly interpreted as the incidence rate ratio. The HR is commonly and conveniently estimated via a Cox proportional hazards model, which can include potential confounders as covariates.

Unfortunately, the use of the HR for causal inference is not straightforward even in the absence of unmeasured confounding, measurement error, and model misspecification. Endowing a HR with a causal interpretation is risky for 2 key reasons: the HR may change over time, and the HR has a built-in selection bias. Here I review these 2 problems and some proposed solutions. As an example, I will use the findings from a Women's Health Initiative randomized experiment that compared the risk of coronary heart disease of women assigned to combined (estrogen plus progestin) hormone therapy with that of women assigned to placebo.¹ By using a randomized experiment as an example, the discussion can focus on the shortcomings of the HR, setting aside issues of confounding and other serious problems that arise in observational studies.

The Women's Health Initiative followed over 16,000 women for an average of 5.2 years before the study was halted due to safety concerns. The primary result from the trial was a HR. As stated in the abstract 1 and shown in Table 1 of the article, "Combined hormone therapy was associated with a hazard ratio of 1.24."¹ In addition, Table 2 provided the HRs during each year of follow-up: 1.81, 1.34, 1.27, 1.25, 1.45, and 0.70 for years 1, 2, 3, 4, 5, and 6 or more, respectively. Thus, the HR reported in the abstract and Table 1 can be viewed as some sort of weighted average of the period-specific HRs reported in Table 2.

Similar articles in PubMed

Hazard ratio bias in cohort studies.

[Epidemiology. 2011]

[Cox regression analysis in epidemiological research].

[G Ital Nefrol. 2011]

Cox proportional hazards models have more statistical power than logistic regression models in cross-sectic [Eur J Hum Genet. 2011]

Regression analysis.

[Pract Neurol. 2011]

Statistical hypothesis testing: associating patient characteristics with an incident condition: K [J Wound Ostomy Continence Nu

See review

See

Cited by other articles in PMC

Risk of tuberculosis in patients with solid cancers and haematological malignancies: [The European Respiratory Journal. 2011]

Is cardiovascular risk reduction therapy effective in South Asian Chinese and other patients with diabetes? A po [BMJ Open. 2011]

Erythrocyte omega-3 fatty acids are inversely associated with incident dementia: Secondary an [Prostaglandins, leukotrienes,

Time-based measures of treatment effect: reassessment of ticagrelor and clopidogrel from the PLATO trial [Open Heart. 2011]

A DAG-based comparison of interventional effect underestimation between composite endpoint [BMC Medical Research Methodolo

See

Links

PubMed

Annals of Internal Medicine

The Spectrum of Subclinical Primary Hypertension

A Cohort Study

Jenifer M. Brown, MD; Cassianne Robinson-Cohen, PhD; Miguel Angel Luque-Fernandez, MSc, MPH, PhD; Matthew A. Allison, MD, MPH; Rene Baudrand, MD; Joachim H. Ix, MD, MS; Bryan Kestenbaum, MD, MS; Ian H. de Boer, MD, MS; and Anand Vaidya, MD, MMSc

Background: Primary aldosteronism is recognized as a severe form of renin-independent aldosteronism that results in excessive mineralocorticoid receptor (MR) activation.

Objective: To investigate whether a spectrum of subclinical renin-independent aldosteronism that increases risk for hypertension exists among normotensive persons.

Design: Cohort study.

Setting: National community-based study.

Participants: 850 untreated normotensive participants in MESA (Multi-Ethnic Study of Atherosclerosis) with measurements of serum aldosterone and plasma renin activity (PRA).

Measurements: Longitudinal analyses investigated whether al-

Editor's comment: RISK DIFFERENCES

"While the findings of the longitudinal component of the analysis are based mostly on **hazard ratios**, the editors also now routinely request that in cohort studies the authors present the findings in a way that provide some understanding of **absolute risks or risk differences**"

(incidence rates per 1000 person-years of follow-up: suppressed renin phenotype, 85.4 events [95% CI, 73.4 to 99.3 events]; indeterminate renin phenotype, 53.3 events [CI, 42.8 to 66.4 events]; unsuppressed renin phenotype, 54.5 events [CI, 41.8 to 71.0 events]). With renin suppression, higher aldosterone concentrations were independently associated with an increased risk for incident hypertension, whereas no association between aldosterone and hypertension was seen when renin was not suppressed. Higher aldosterone concentrations were associated with lower serum potassium and higher urinary excretion of potassium, but only when renin was suppressed.

Limitation: Sodium and potassium were measured several years before renin and aldosterone.

Conclusion: Suppression of renin and higher aldosterone con-

Annals of Internal Medicine

The Spectrum of Subclinical Primary Hypertension

A Cohort Study

Jenifer M. Brown, MD; Cassianne Robinson-Cohen, PhD; Miguel A. Allison, MD, MPH; Rene Baudrand, MD; Joachim F. and Anand Vaidya, MD, MMSc

Background: Primary aldosteronism is recognized as a severe form of renin-independent aldosteronism that results in excessive mineralocorticoid receptor (MR) activation.

Objective: To investigate whether a spectrum of subclinical renin-independent aldosteronism that increases risk for hypertension exists among normotensive persons.

Design: Cohort study.

Setting: National community-based study.

Participants: 850 untreated normotensive participants in MESA (Multi-Ethnic Study of Atherosclerosis) with measurements of serum aldosterone and plasma renin activity (PRA).

Measurements: Longitudinal analyses investigated whether al-

Editor's comment: RISK DIFFERENCES

For instance, by presenting **adjusted survival curves** and 5-year (or 8-year) **adjusted cumulative incidence** of hypertension, with either **risk ratios** or **differences**, by category of plasma renin activity and/or aldosterone levels. You can find an example of this approach in the paper by **Chang et al in Ann Intern Med 2016;164(5):305-12**, although there are several valid approaches to this problem. We believe that this presentation provides a better understanding of the association between exposure and outcomes than just presenting of hazard ratios.

aldosterone and hypertension was seen when renin was not suppressed. Higher aldosterone concentrations were associated with lower serum potassium and higher urinary excretion of potassium, but only when renin was suppressed.

Limitation: Sodium and potassium were measured several years before renin and aldosterone.

Conclusion: Suppression of renin and higher aldosterone con-

CER is about Causal Inference

Annals of Internal Medicine

ORIGINAL RESEARCH

Metabolically Healthy Obesity and Development of Chronic Kidney Disease

A Cohort Study

Yi-Jin Choi, MD, PhD; Seung-Ho Ryu, MD, PhD; Yest Choi, BS; Yul Dang, PhD; Jahn Cho, PhD; Min-Jaeo Kwon, MD, PhD; Young Goo Hyun, MD, PhD; Kyu-Bock Lee, MD, PhD; Hyun-Kyu Kim, MD, PhD; Hyun-Sik Jung, MD; Kyung Sun Yoo, MD, PhD; Jit Ahn, MSc; Seojay Bang, MD, PhD; Di Zhao, PhD; Hyang-Seung Soh, MD, PhD; Sun Cheol Chung, MD, PhD; Hocheol Shin, MD, PhD; Roberto Pastor-Barriuso, PhD; and Illeses Guillar, MD, DrPH

Background: The risk for chronic kidney disease (CKD) among obese persons without obesity-related metabolic abnormalities, called metabolically healthy obesity, is largely unexplored.

Objective: To investigate the risk for incident CKD across categories of body mass index in a large cohort of metabolically healthy men and women.

Design: Prospective cohort study.

Setting: Kangbuk Samsung Health Study, Kangbuk Samsung Hospital, Seoul, South Korea.

Participants: 62 249 metabolically healthy, young and middle-aged men and women without CKD or proteinuria at baseline.

Measurements: Metabolically healthy was defined as a homogeneous group of individuals with a body mass index of 25 and absence of any component of the metabolic syndrome. Underweight, normal weight, overweight, and obesity were defined as a body mass index less than 18.5 kg/m², 18.5 to 22.9 kg/m², 23 to 24.9 kg/m², and 25 kg/m² or greater, respectively. Incident CKD was incident CKD, defined as an estimated glomerular filtration rate less than 60 mL/min/1.73 m².

Chronic kidney disease (CKD) is a major clinical condition and public health problem (1). It is a precursor for cardiovascular diseases and mortality (2). Its prevalence is increasing worldwide along with the growing prevalence of obesity and metabolic disease (3). Indeed, obesity-mediated by hypertension, insulin resistance, dyslipidemia, and reduced physical activity, metabolic abnormalities is a major risk factor for CKD (4).

Although the role of obesity-induced metabolic abnormalities in CKD development is well established, metabolically healthy obese (MHO) persons seem to have a favorable profile with no metabolic abnormalities (5, 6). The association between MHO and CKD, however, remains unclear. The previous studies have found no association (7), but the comparison between MHO and normal-weight participants could be biased because the reference group included over-weight individuals. In our study, all the participants were defined as those with fewer than 2 metabolic components. Therefore, we examined the association between categories of body mass index and CKD in a large sample of metabolically healthy men and women who had health screening examinations.

Results: During 349 028 person-years of follow-up, 906 incident CKD cases were identified. The multivariable-adjusted differences in 5-year cumulative incidence of CKD in underweight, overweight, and obese participants compared with metabolically healthy participants were 4.0% (1.8%–6.5%; IC, 0.4 to 8.1), and 6.7% (3.0 to 10.4) cases per 1000 persons, respectively. These associations were consistently seen in all clinically relevant subgroups.

Limitations: Chronic kidney disease was identified by a single measurement.

Conclusion: Overweight and obesity are associated with an increased incidence of CKD in metabolically healthy young and middle-aged participants. These findings show that metabolically healthy obesity is not a harmless condition and that the categories of obesity, regardless of metabolic abnormalities, can adversely affect renal function.

Primary Funding Source: None.

Ann Intern Med. 2016;164:605–612. doi:10.7326/M15-1323 www.annals.org

For author affiliations, see page 606.

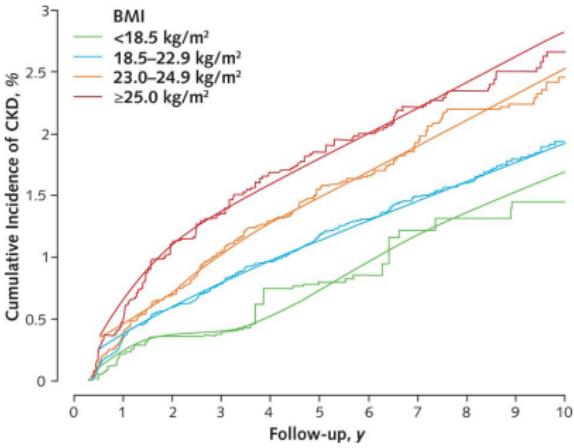
This article was published at www.annals.org on 9 February 2016.

Methods

Study Population

The Kangbuk Samsung Health Study is a cohort study of Korean men and women aged 18 years or older who had a comprehensive annual or biennial health examination at the clinics of the Kangbuk Samsung Hospital Health Screening Centers in Seoul and Suwon, South Korea (8). More than 80% of participants were employees of various organizations and local government organizations and their spouses. In South Korea, the Industrial Safety and Health Act requires all employees to receive annual or biennial health screening examinations, offered free of charge. The remaining participants gathered for the screening examinations on their own.

Our analysis included all persons who had comprehensive health examinations from 1 January 2002 to 31 December 2009 and had at least one follow-up examination before 31 December 2013 (that is, those who had a baseline visit and at least one follow-up visit [$n = 175,859$] (Figure 1). We excluded persons who had metabolic abnormalities (5, 9, 10) and evidence of kidney disease at baseline, such as proteinuria. We also excluded persons whose serum creatinine levels were greater than 1.5 mg/dL or whose serum glucose levels of 100 mg/dL or greater or who used glucose-lowering agents, blood pressure (BP) of



Rubin and Heckman

- This framework was developed first by statisticians (Rubin, 1983) and econometricians (Heckman, 1978) as a new approach for the estimation of **causal effects** from observational data.
- We will keep separate the **causal framework** (a conceptual issue briefly introduce here) and the "**how to estimate causal effects**" (an statistical issue also introduced here)

Notation and definitions

Observed Data

- Treatment **A**.

Often, $A = 1$ for treated and $A = 0$ for control.

- Confounders **W**.
- Outcome **Y**.

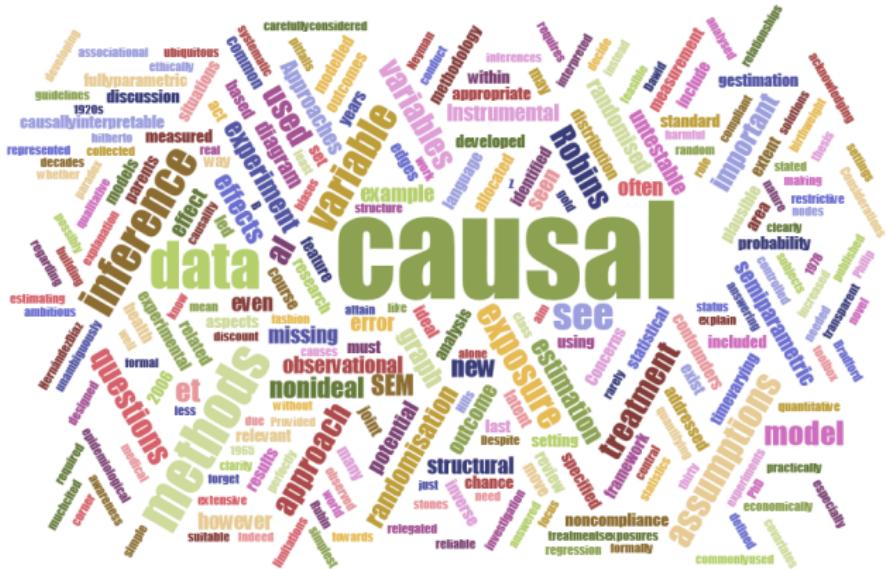
Potential Outcomes

- For patient i $\mathbf{Y}_i(1)$ and $\mathbf{Y}_i(0)$ set to $A = a$ $\mathbf{Y}^{(a)}$, namely $A = 1$ and $A = 0$.

Causal Effects

- Average Treatment Effect: $E[\mathbf{Y}(1) - \mathbf{Y}(0)]$.

COUNTERFACTUAL FRAMEWORK



When an RCT is not possible

- This framework was developed first by statisticians (Rubin, 1983) and econometricians (Heckman, 1978) as a new approach for the estimation of **causal effects** from observational data.

When an RCT is not possible

- This framework was developed first by statisticians (Rubin, 1983) and econometricians (Heckman, 1978) as a new approach for the estimation of **causal effects** from observational data.
- Classically known as the **Neyman-Rubin Counterfactual Framework**.

When an RCT is not possible

- This framework was developed first by statisticians (Rubin, 1983) and econometricians (Heckman, 1978) as a new approach for the estimation of **causal effects** from observational data.
- Classically known as the **Neyman-Rubin Counterfactual Framework**.
- The counterfactual framework offers an approach to **IE** when a Random Clinical Trial (**RCT**) is **unfeasible or unethical**.

When an RCT is not possible

- This framework was developed first by statisticians (Rubin, 1983) and econometricians (Heckman, 1978) as a new approach for the estimation of **causal effects** from observational data.
- Classically known as the **Neyman-Rubin Counterfactual Framework**.
- The counterfactual framework offers an approach to **IE** when a Random Clinical Trial (**RCT**) is **unfeasible or unethical**.

When a RCT is not possible

- The counterfactual framework offers an approach to IE (effectiveness) when researchers need to assess treatment effects from survey data, census data, administrative data, or other types of data.

When a RCT is not possible

- The counterfactual framework offers an approach to IE (effectiveness) when researchers need to assess treatment effects from survey data, census data, administrative data, or other types of data.
- ***"Data collected through the observation of systems as they operate in normal practice without any interventions implemented by randomized assignments rules"*** (Rubin, 1977, p.757)

When a RCT is not possible

- The counterfactual framework offers an approach to IE (effectiveness) when researchers need to assess treatment effects from survey data, census data, administrative data, or other types of data.
- ***"Data collected through the observation of systems as they operate in normal practice without any interventions implemented by randomized assignments rules"*** (Rubin, 1977, p.757)
- Big opportunity for IE in the era of the **BIG DATA REVOLUTION** (e.g., Digital medical records, Births cohorts, international HIV cohorts, <http://www.iedea-sa.org/>, Biobanks UK-Biobank, etcetera.)

Counterfactual

The main challenge across different types of **IE** is to find a **good counterfactual** -namely, *the situation a participating subject would have experienced had he or she not been exposed to the program or intervention.*

Causal effects in the real world

$$\text{ATE} = [E(Y_i(1) | A = 1)] - [E(Y_i(0) | A = 0)]$$

Causal effects in the real world

$$\text{ATE} = [E(Y_i(1) | A = 1)] - [E(Y_i(0) | A = 0)]$$

The **Outcomes** for the treated and control individuals are:

Causal effect

Causal effects in the real world

$$\text{ATE} = [E(Y_i(1) | A = 1)] - [E(Y_i(0) | A = 0)]$$

The **Outcomes** for the treated and control individuals are:

$Y_i(1) = Y_i(A = 1)$ for some treatment variable A (**Treated**).

Causal effects in the real world

$$\text{ATE} = [E(Y_i(1) | A = 1)] - [E(Y_i(0) | A = 0)]$$

The **Outcomes** for the treated and control individuals are:

$Y_i(1) = Y_i(A = 1)$ for some treatment variable A (**Treated**).

$Y_i(0) = Y_i(A = 0)$ for some treatment variable A (**Control**).

Causal effects in the real world

$$\text{ATE} = [E(Y_i(1) | A = 1)] - [E(Y_i(0) | A = 0)]$$

The **Outcomes** for the treated and control individuals are:

$Y_i(1) = Y_i(A = 1)$ for some treatment variable A (**Treated**).

$Y_i(0) = Y_i(A = 0)$ for some treatment variable A (**Control**).

Causal effects in an ideal world

The **Potential Outcomes** for an individual i if he/she received treatment or control are:

Causal effects in an ideal world

The **Potential Outcomes** for an individual i if he/she received treatment or control are:

$Y_i(0) = Y_i(A = 0)$ is the counterfactual or potential outcome for
 $Y_i(0) = Y_i(A = 1)$ (**Treated**)

Causal effects in an ideal world

The **Potential Outcomes** for an individual i if he/she received treatment or control are:

$Y_i(0) = Y_i(A = 0)$ is the counterfactual or potential outcome for
 $Y_i(0) = Y_i(A = 1)$ (**Treated**)

$Y_i(1) = Y_i(A = 1)$ is the counterfactual or potential outcome for
 $Y_i(1) = Y_i(A = 0)$ (**Control**)

Causal effects in an ideal world

The **Potential Outcomes** for an individual i if he/she received treatment or control are:

$Y_i(0) = Y_i(A = 0)$ is the counterfactual or potential outcome for
 $Y_i(0) = Y_i(A = 1)$ (**Treated**)

$Y_i(1) = Y_i(A = 1)$ is the counterfactual or potential outcome for
 $Y_i(1) = Y_i(A = 0)$ (**Control**)

However, we only observe:

Causal effect

Causal effects in an ideal world

The **Potential Outcomes** for an individual i if he/she received treatment or control are:

$Y_i(0) = Y_i(A = 0)$ is the counterfactual or potential outcome for
 $Y_i(0) = Y_i(A = 1)$ (**Treated**)

$Y_i(1) = Y_i(A = 1)$ is the counterfactual or potential outcome for
 $Y_i(1) = Y_i(A = 0)$ (**Control**)

However, we only observe:

$Y_i(1) = Y_i(A = 1)$ and $Y_i(0) = Y_i(A = 0)$

The fundamental problem of Causal inference

The counterfactual is **not observed**.

So the challenge of an **IE** is to create a convincing and reasonable comparison group for beneficiaries in light of this **missing data**.

Causal effect

The fundamental problem of Causal inference

The counterfactual is **not observed**.

So the challenge of an **IE** is to create a convincing and reasonable comparison group for beneficiaries in light of this **missing data**.

Total Causal Effect

$$[(Y_i(1) \mid A = 1) + (\text{PO})] - [(Y_i(0) \mid A = 0) + (\text{PO})]$$

The fundamental problem of Causal inference

The counterfactual is **not observed**.

So the challenge of an **IE** is to create a convincing and reasonable comparison group for beneficiaries in light of this **missing data**.

Total Causal Effect

$$[(Y_i(1) \mid A = 1) + (Y_i(1) \mid A = 0)] - [(Y_i(0) \mid A = 0) + (Y_i(0) \mid A = 1)]$$

Causal effect

The fundamental problem of Causal inference

The counterfactual is **not observed**.

So the challenge of an **IE** is to create a convincing and reasonable comparison group for beneficiaries in light of this **missing data**.

The fundamental problem of Causal inference: selection bias (confounding)

$$[(Y_i(1) \mid A = 1) + (Y_i(1) \mid A = 0)] - [(Y_i(0) \mid A = 0) + (Y_i(0) \mid A = 1)]$$

Causal effect

The fundamental problem of Causal inference

The counterfactual is **not observed**.

So the challenge of an **IE** is to create a convincing and reasonable comparison group for beneficiaries in light of this **missing data**.

The fundamental problem of Causal inference: selection bias (confounding)

$$[(Y_i(1) \mid A = 1) + (Y_i(1) \mid A = 0)] - [(Y_i(0) \mid A = 0) + (Y_i(0) \mid A = 1)]$$

Causal effect in OBSERVATIONAL study

ASSUMPTIONS to consider a CAUSAL EFFECT

- Rosebaum & Rubin, 1983: **The Ignorable Treatment Assignment Assumption** (Unconfoundedness, conditional mean Independence, ignorability or exchangeability).

$$(Y_i(1), Y_i(0)) \perp A_i \mid W_i$$

Causal effect in an EXPERIMENTAL study

The solution to the fundamental problem of Causal inference

$$[(Y_i | A = 1) + (Y_i | A = 0)] - [(Y_i | A = 0) + (Y_i | A = 1)]$$

Causal effect in an EXPERIMENTAL study

The solution to the fundamental problem of Causal inference

$$[(Y_i | A = 1) + (Y_i | A = 0)] - [(Y_i | A = 0) + (Y_i | A = 1)]$$

Causal effect in an EXPERIMENTAL study

The solution to the fundamental problem of Causal inference

$$[(Y_i | A = 1) + (Y_i | A = 0)] - [(Y_i | A = 0) + (Y_i | A = 1)]$$

RANDOMIZATION makes the ATE UNBIASED:

$$[(Y_i | A = 1) = (Y_i | A = 0)] - [(Y_i | A = 0) = (Y_i | A = 1)]$$

Causal effect in an EXPERIMENTAL study

The solution to the fundamental problem of Causal inference

$$[(Y_i | A = 1) + (Y_i | A = 0)] - [(Y_i | A = 0) + (Y_i | A = 1)]$$

RANDOMIZATION makes the ATE UNBIASED:

$$[(Y_i | A = 1) = (Y_i | A = 0)] - [(Y_i | A = 0) = (Y_i | A = 1)]$$

Causal effect in an EXPERIMENTAL study

The solution to the fundamental problem of Causal inference

$$[(Y_i | A = 1) + (Y_i | A = 0)] - [(Y_i | A = 0) + (Y_i | A = 1)]$$

RANDOMIZATION makes the **ATE** UNBIASED:

$$[(Y_i | A = 1) = (Y_i | A = 0)] - [(Y_i | A = 0) = (Y_i | A = 1)]$$

The solution to the fundamental problem of Causal inference

$$\text{ATE} = [E(Y_i | A = 1)] - [E(Y_i | A = 0)]$$

Causal effects in OBSERVATIONAL studies

When randomization is unethical or infeasible

Causal effect is biased (B):

$$\text{ATE} + \mathbf{B}$$

Type of bias

- ① **Observed:** The treatment assignment is not random.

Causal effects in OBSERVATIONAL studies

When randomization is unethical or infeasible

Causal effect is biased (B):

$$\text{ATE} + \mathbf{B}$$

Type of bias

- ① **Observed:** The treatment assignment is not random.
- ② **Unobserved:** Unobserved factors associated with both the treatment and the effect.

Causal effects in OBSERVATIONAL studies

When randomization is unethical or infeasible

Causal effect is biased (B):

$$\text{ATE} + \mathbf{B}$$

Type of bias

- ① **Observed:** The treatment assignment is not random.
- ② **Unobserved:** Unobserved factors associated with both the treatment and the effect.

ASSUMPTIONS to consider a CAUSAL EFFECT

- Rosebaum & Rubin, 1983: **The Ignorable Treatment Assignment Assumption** (Unconfoundedness, conditional mean Independence or ignorability).
 $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$

ASSUMPTIONS to consider a CAUSAL EFFECT

- Rosebaum & Rubin, 1983: **The Ignorable Treatment Assignment Assumption** (Unconfoundedness, conditional mean Independence or ignorability).
 $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$
- IID random variables.

ASSUMPTIONS to consider a CAUSAL EFFECT

- Rosebaum & Rubin, 1983: **The Ignorable Treatment Assignment Assumption** (Unconfoundedness, conditional mean Independence or ignorability).
 $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$
- **IID** random variables.
- The model used to estimate the assignment probability has to **be correctly specified**.

ASSUMPTIONS to consider a CAUSAL EFFECT

- Rosebaum & Rubin, 1983: **The Ignorable Treatment Assignment Assumption** (Unconfoundedness, conditional mean Independence or ignorability).
 $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$
- **IID** random variables.
- The model used to estimate the assignment probability has to **be correctly specified**.
- **POSITIVITY**: We assume $P(A = a \mid W) > 0$ for all a, W .

ASSUMPTIONS to consider a CAUSAL EFFECT

- Rosebaum & Rubin, 1983: **The Ignorable Treatment Assignment Assumption** (Unconfoundedness, conditional mean Independence or ignorability).
 $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$
- **IID** random variables.
- The model used to estimate the assignment probability has to **be correctly specified**.
- **POSITIVITY**: We assume $P(A = a \mid W) > 0$ for all a, W .
- **SUTVA**: We assume that there is **only one version of the treatment**: **CONSISTENCY** and the assignment to the treatment to one unit doesn't affect the outcome of another unit.

ASSUMPTIONS to consider a CAUSAL EFFECT

- Rosebaum & Rubin, 1983: **The Ignorable Treatment Assignment Assumption** (Unconfoundedness, conditional mean Independence or ignorability).
 $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$
- **IID** random variables.
- The model used to estimate the assignment probability has to **be correctly specified**.
- **POSITIVITY**: We assume $P(A = a \mid W) > 0$ for all a, W .
- **SUTVA**: We assume that there is **only one version of the treatment**: **CONSISTENCY** and the assignment to the treatment to one unit doesn't affect the outcome of another unit.

How to use the Neyman-Rubin Counterfactual framework?

- ① Be guided by the appropriate theory in our area of research.

How to use the Neyman-Rubin Counterfactual framework?

- ① Be guided by the appropriate theory in our area of research.
- ② We have to have a substantial knowledge of the context and program to evaluate.

How to use the Neyman-Rubin Counterfactual framework?

- ① Be guided by the appropriate theory in our area of research.
- ② We have to have a substantial knowledge of the context and program to evaluate.
- ③ Apply the Neyman-Rubin Counterfactual Framework in order to consider an effect to be CAUSAL.

How to use the Neyman-Rubin Counterfactual framework?

- ① Be guided by the appropriate theory in our area of research.
- ② We have to have a substantial knowledge of the context and program to evaluate.
- ③ Apply the Neyman-Rubin Counterfactual Framework in order to consider an effect to be CAUSAL.

The best design

The gold standard of scientific research

- ① The simplest way to conceptually compare treated and control units.

The best design

The gold standard of scientific research

- ① The simplest way to conceptually compare treated and control units.
- ② Units are randomly assigned to treatment and control groups.

The best design

The gold standard of scientific research

- ① The simplest way to conceptually compare treated and control units.
- ② Units are randomly assigned to treatment and control groups.
- ③ If randomization is not possible, use the conditional ignorability assumption.

The best design

The gold standard of scientific research

- ① The simplest way to conceptually compare treated and control units.
- ② Units are randomly assigned to treatment and control groups.
- ③ If randomization is not possible, use the conditional ignorability assumption.
- ④ If experimentation is not possible, try to mimic the principles of experimentation in your study design.

The best design

The gold standard of scientific research

- ① The simplest way to conceptually compare treated and control units.
- ② Units are randomly assigned to treatment and control groups.
- ③ If randomization is not possible, use the conditional ignorability assumption.
- ④ If experimentation is not possible, try to mimic the principles of experimentation in your study design.
- ⑤ In both quasi-experimental and experimental designs use the intention to treat approach.

The best design

The gold standard of scientific research

- ① The simplest way to conceptually compare treated and control units.
- ② Units are randomly assigned to treatment and control groups.
- ③ If randomization is not possible, use the conditional ignorability assumption.
- ④ If experimentation is not possible, try to mimic the principles of experimentation in your study design.
- ⑤ In both quasi-experimental and experimental designs use the intention to treat approach.

Causal effect

The fundamental problem of Causal inference

In the real world, this is what we see:

$Unit_i$	W_i^1	W_i^2	W_i^3	A_i	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
1	2	1	503	0	693	?	?
2	7	1	985	0	111	?	?
3	8	2	830	1	?	102	?
4	3	1	938	1	?	111	?

Causal effect

The fundamental problem of Causal inference

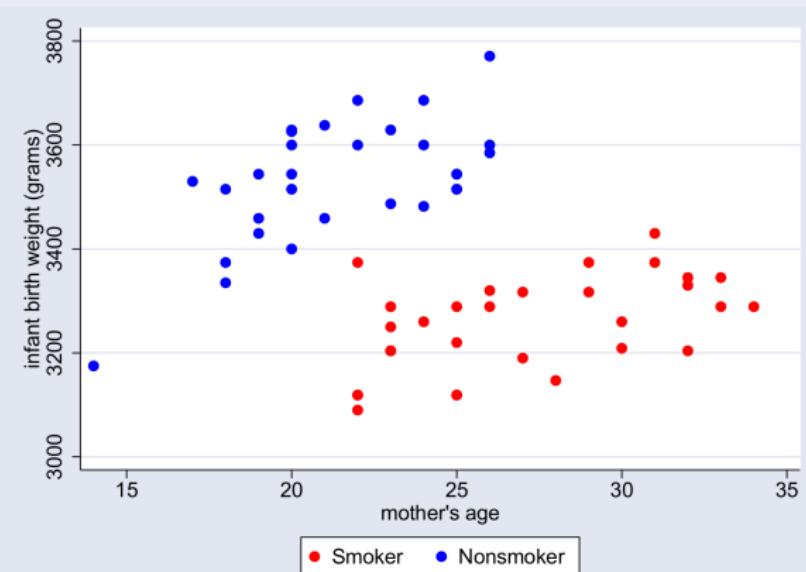
In an ideal world, we would see this:

$Unit_i$	W_i^1	W_i^2	W_i^3	A_i	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
1	2	1	503	0	693	75	-618
2	7	1	985	0	111	108	-3
3	8	2	830	1	944	102	-842
4	3	1	938	1	14	111	97

Causal effect

The fundamental problem of Causal inference

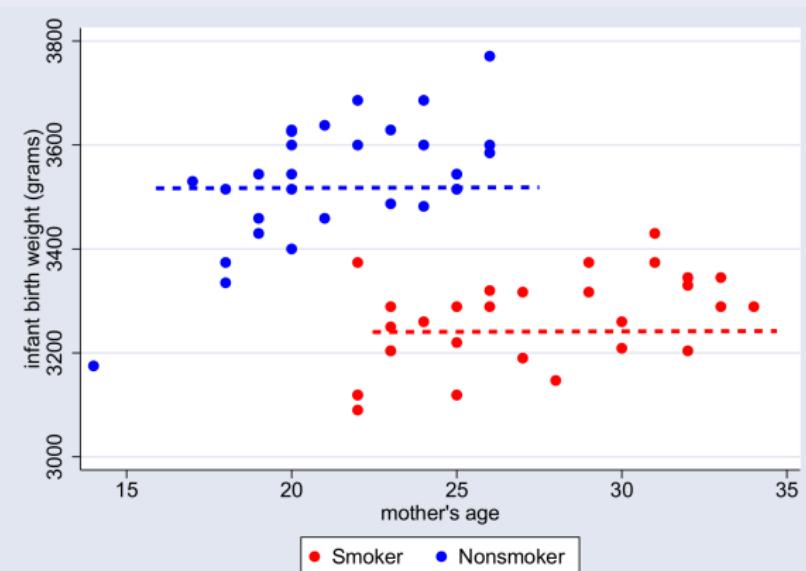
In the real world, this is what we see:



Causal effect

The fundamental problem of Causal inference

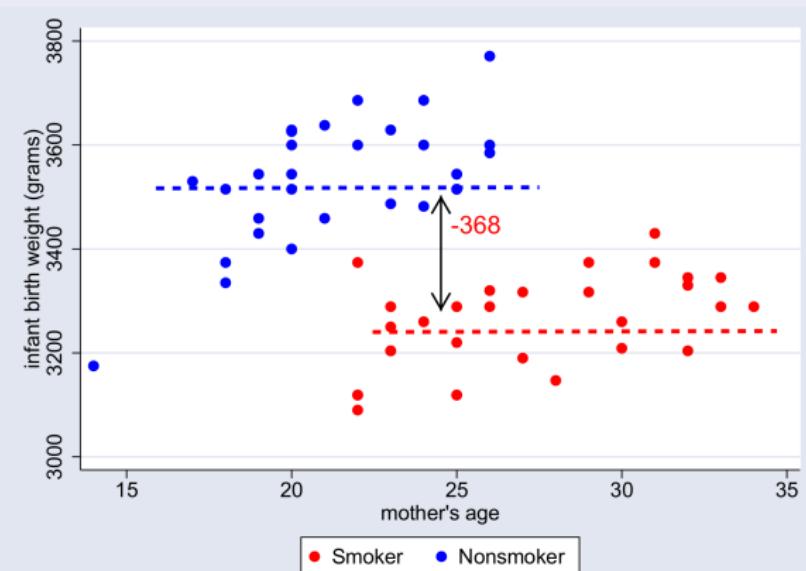
In the real world, this is what we see:



Causal effect

The fundamental problem of Causal inference

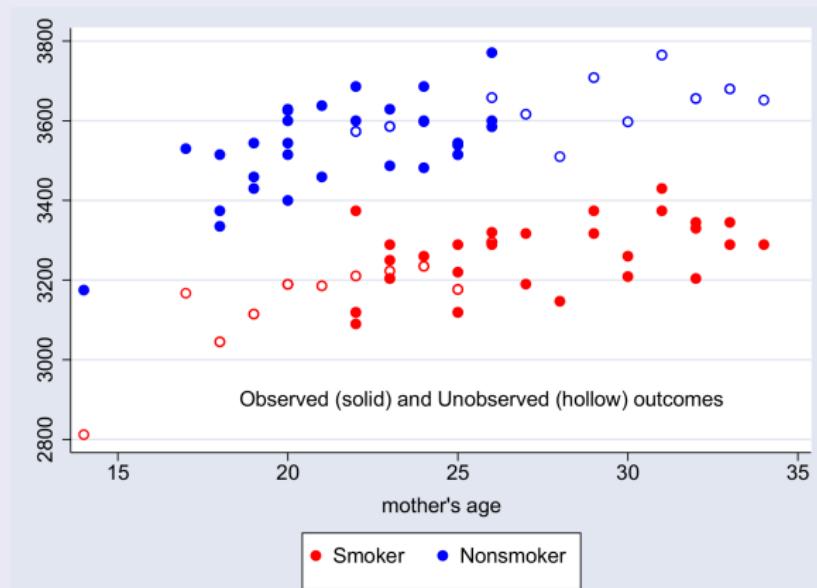
In the real world, this is what we see:



Causal effect

The fundamental problem of Causal inference

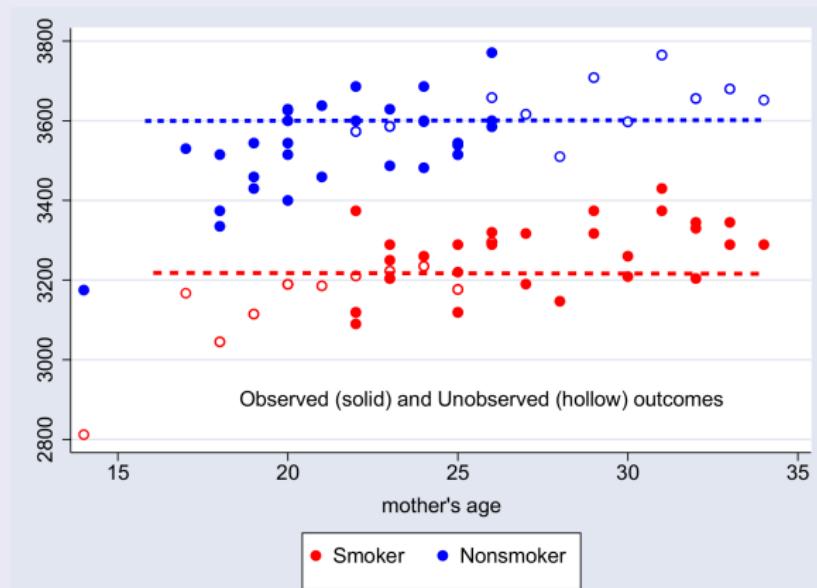
In an ideal world, this is what we see:



Causal effect

The fundamental problem of Causal inference

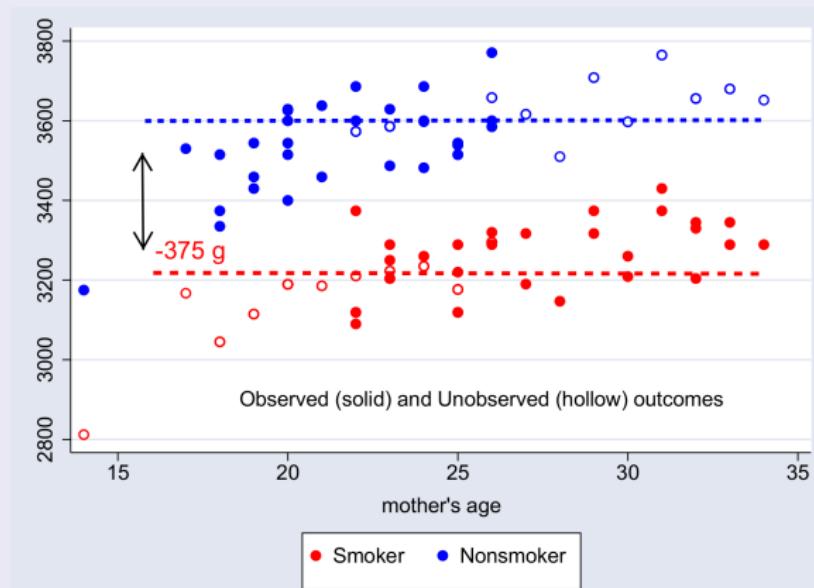
In an ideal world, this is what we see:



Causal effect

The fundamental problem of Causal inference

In an ideal world, this is what we see:



Review: Causal effects with observational data

Potential Outcomes

Treatment (A) effect on outcome (Y) in real world:

$$Y_i(1) = Y_i(A = 1) \text{ and } Y_i(0) = Y_i(A = 0)$$

However we would like to know what would have happened if:

Treated $Y_i(1)$ would have been non-treated $Y_i(A = 0) = Y_i(0)$.

Controls $Y_i(0)$ would have been treated $Y_i(A = 1) = Y_i(1)$.

Identifiability

- How we can identify the effect of the potential outcomes Y^a if they are not observed?
- How we can estimate the expected difference between the potential outcomes $E[Y(1) - Y(0)]$, namely the **ATE**.

Background: Causal Inference Assumptions

IGNORABILITY

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i | W_i$$

POSITIVITY

POSITIVITY: $P(A = a | W) > 0$ for all a, W

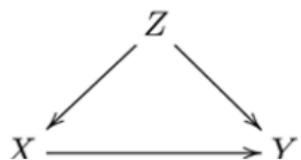
SUTVA

- We have assumed that there is **only one version of the treatment (consistency)** $Y(1)$ if $A = 1$ and $Y(0)$ if $A = 0$ (i.e., $Y = AY(1) + (1-A)Y(0)$).
- The assignment to the treatment to one unit doesn't affect the outcome of another unit (no **interference** or **IID** random variables).
- Assignment probability model has to be **correctly specified**.

Causal Inference Potential Outcomes IDENTIFIABILITY and IGNORABILITY

We use DAGs to evaluate identifiability and ignorability

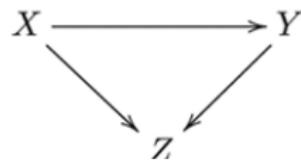
A



B



C



Better control of confounding: Ignorability

- Ignorability refers to the treatment assignment being independent of potential outcomes conditional on some set of confounders W :

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$$

Better control of confounding: Ignorability

- Ignorability refers to the treatment assignment being independent of potential outcomes conditional on some set of confounders W :

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$$

- Ignorability assumes that treatment assignment is **random** given W (i.e., a collection of variables).

Better control of confounding: Ignorability

- Ignorability refers to the treatment assignment being independent of potential outcomes conditional on some set of confounders W :

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$$

- Ignorability assumes that treatment assignment is **random** given W (i.e., a collection of variables).
- So the question is **what collection** of variables?

Better control of confounding: Ignorability

- Ignorability refers to the treatment assignment being independent of potential outcomes conditional on some set of confounders W :

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$$

- Ignorability assumes that treatment assignment is **random** given W (i.e., a collection of variables).
- So the question is **what collection** of variables?
- We need to identify what a set of variables like this, then that's sufficient to **control for confounding**.

Better control of confounding: Ignorability

- Ignorability refers to the treatment assignment being independent of potential outcomes conditional on some set of confounders W :

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$$

- Ignorability assumes that treatment assignment is **random** given W (i.e., a collection of variables).
- So the question is **what collection** of variables?
- We need to identify what a set of variables like this, then that's sufficient to **control for confounding**.
- **DAGs** help to find **W** , based on subject matter knowledge, that will **achieve** ignorability (A.K.A. exchangeability).

Better control of confounding: Ignorability

- Ignorability refers to the treatment assignment being independent of potential outcomes conditional on some set of confounders W :

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i \mid W_i$$

- Ignorability assumes that treatment assignment is **random** given W (i.e., a collection of variables).
- So the question is **what collection** of variables?
- We need to identify what a set of variables like this, then that's sufficient to **control for confounding**.
- **DAGs** help to find **W** , based on subject matter knowledge, that will **achieve** ignorability (A.K.A. exchangeability).

Nomenclature

- A **Directed Acyclic Graph** (DAG) is a causal diagram where all arrows are directed and represent causal effects on one variable or another, and is acyclic in that one cannot return to where one started via directed arrows.

Nomenclature

- A **Directed Acyclic Graph** (DAG) is a causal diagram where all arrows are directed and represent causal effects on one variable or another, and is acyclic in that one cannot return to where one started via directed arrows.
- **Back door path:** A non-causal path in a DAG from exposure (A) to outcome (Y) that has an arrow coming into the exposure (A) and other to the outcome (Y). If there is no collider on the back door path, it is open and requires blocking by conditioning for one of more variables on the path ($Z < - W - > Y$).

Nomenclature

- A **Directed Acyclic Graph** (DAG) is a causal diagram where all arrows are directed and represent causal effects on one variable or another, and is acyclic in that one cannot return to where one started via directed arrows.
- **Back door path:** A non-causal path in a DAG from exposure (A) to outcome (Y) that has an arrow coming into the exposure (A) and other to the outcome (Y). If there is no collider on the back door path, it is open and requires blocking by conditioning for one or more variables on the path ($Z < - W - > Y$).
- **Front door path:** A causal path in a DAG from exposure (A) to outcome (Y) that has an arrow going out of exposure, and arrow into the outcome, and no colliders.

Nomenclature

- A **Directed Acyclic Graph** (DAG) is a causal diagram where all arrows are directed and represent causal effects on one variable or another, and is acyclic in that one cannot return to where one started via directed arrows.
- **Back door path:** A non-causal path in a DAG from exposure (A) to outcome (Y) that has an arrow coming into the exposure (A) and other to the outcome (Y). If there is no collider on the back door path, it is open and requires blocking by conditioning for one or more variables on the path ($Z < - W - > Y$).
- **Front door path:** A causal path in a DAG from exposure (A) to outcome (Y) that has an arrow going out of exposure, and arrow into the outcome, and no colliders.

Properties of a confounder

- Must be associated with the exposure (A) in the source population.

Properties of a confounder

- Must be associated with the exposure (A) in the source population.
- Must be an extraneous risk factor for the disease (Y).

Properties of a confounder

- Must be associated with the exposure (A) in the source population.
- Must be an extraneous risk factor for the disease (Y).
 - Need not be actual cause of disease (Y), but must be surrogate of cause.

Confounding

Properties of a confounder

- Must be associated with the exposure (A) in the source population.
- Must be an extraneous risk factor for the disease (Y).
 - Need not be actual cause of disease (Y), but must be surrogate of cause.
 - Must be risk factor among non-exposed (in the source population): **positivity**.

Properties of a confounder

- Must be associated with the exposure (A) in the source population.
- Must be an extraneous risk factor for the disease (Y).
 - Need not be actual cause of disease (Y), but must be surrogate of cause.
 - Must be risk factor among non-exposed (in the source population): **positivity**.
- Must be not affected by (common cause of) the exposure or (common cause of disease): **collider**.

Properties of a confounder

- Must be associated with the exposure (A) in the source population.
- Must be an extraneous risk factor for the disease (Y).
 - Need not be actual cause of disease (Y), but must be surrogate of cause.
 - Must be risk factor among non-exposed (in the source population): **positivity**.
- Must be not affected by (common cause of) the exposure or (common cause of disease): **collider**.
- Cannot be an intermediary cause: **mediator**

Properties of a confounder

- Must be associated with the exposure (A) in the source population.
- Must be an extraneous risk factor for the disease (Y).
 - Need not be actual cause of disease (Y), but must be surrogate of cause.
 - Must be risk factor among non-exposed (in the source population): **positivity**.
- Must be not affected by (common cause of) the exposure or (common cause of disease): **collider**.
- Cannot be an intermediary cause: **mediator**

Colliders

- A **collider** for a certain pair of variables (e.g., an outcome Y and an exposure A) is a third variable (C) that is caused by both.

Colliders

- A **collider** for a certain pair of variables (e.g., an outcome Y and an exposure A) is a third variable (C) that is caused by both.
- In a **directed acyclic graph** (DAG), a collider is the variable in the middle of an inverted fork (i.e., the variable C in $A \rightarrow C \leftarrow Y$).

Colliders

- A **collider** for a certain pair of variables (e.g., an outcome Y and an exposure A) is a third variable (C) that is caused by both.
- In a **directed acyclic graph** (DAG), a collider is the variable in the middle of an inverted fork (i.e., the variable C in $A \rightarrow C \leftarrow Y$).

Colliders

- Controlling for, or conditioning an analysis on a collider (i.e., through stratification or regression) can introduce a **spurious association** between its causes.

Colliders

- Controlling for, or conditioning an analysis on a collider (i.e., through stratification or regression) can introduce a **spurious association** between its causes.
- This potentially explains many **paradoxical findings** in the medical literature, where established risk factors for a particular outcome appear protective.

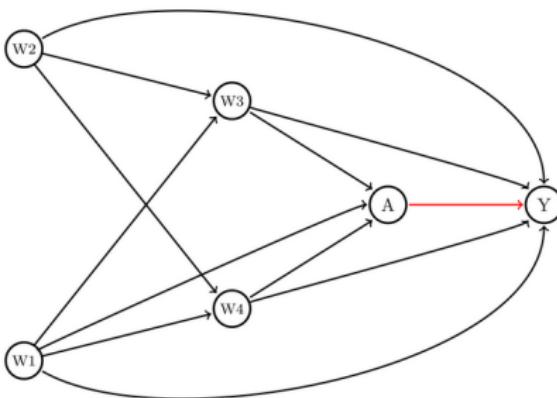
Colliders

- Controlling for, or conditioning an analysis on a collider (i.e., through stratification or regression) can introduce a **spurious association** between its causes.
- This potentially explains many **paradoxical findings** in the medical literature, where established risk factors for a particular outcome appear protective.
- Deconstructing paradoxical effects in medical literature:
Luque-Fernandez MA et al. Deconstructing the **smoking-preeclampsia paradox** through a counterfactual framework. Eur J Epidemiol. 2016;31:613-623
(<https://www.ncbi.nlm.nih.gov/pubmed/26975379>).

DAG example: back door paths

Direct Acyclic Graph (DAG)

Under conditional exchangeability: $Y(0), Y(1) \perp\!\!\!\perp A|W$
ATE = $E[E(Y|A = 1; W) - E(Y|A = 0; W)]$



Y = Mortality; A = Chemotherapy vs. Chemotherapy & Radiotherapy; W_1 = Sex; W_2 = Age; W_3 = TNM-Stage; W_4 = Comorbidities

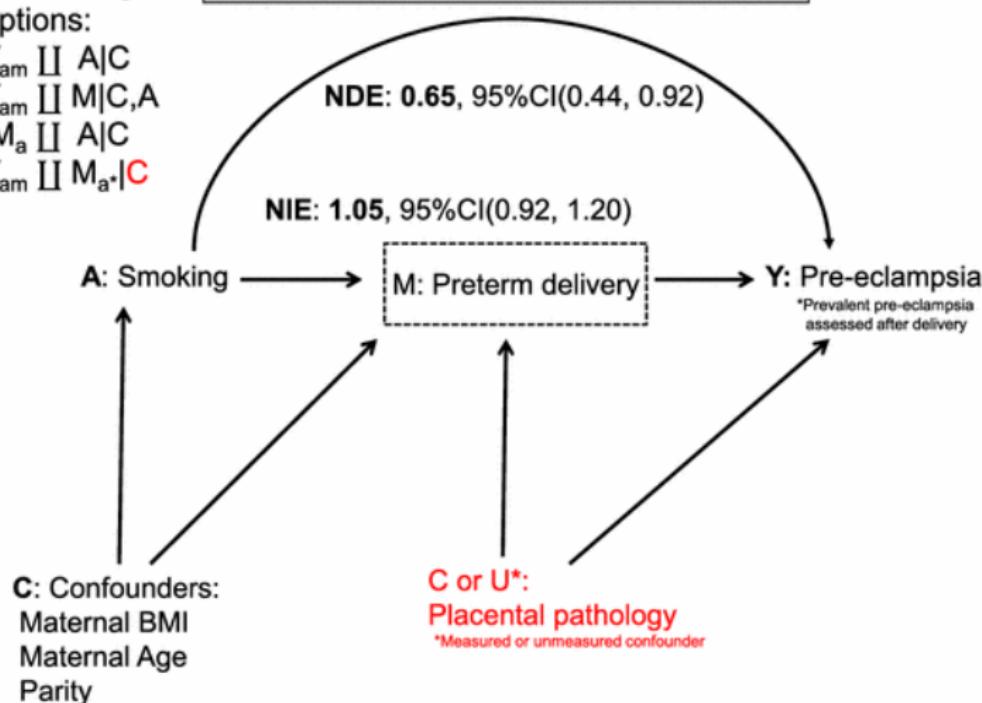
⁵ Luque-Fernandez MA et al. Data-Adaptive Estimation for Double-Robust Methods in Population-Based Cancer Epidemiology: Risk differences for lung cancer mortality by emergency presentation (2017). AJE.
<https://academic.oup.com/aje/article/doi/10.1093/aje/kwx317/4110407>

DAG example: mediation

Mediation analysis: Marginal Total Effect: **0.68**, 95%CI(0.45, 0.97)

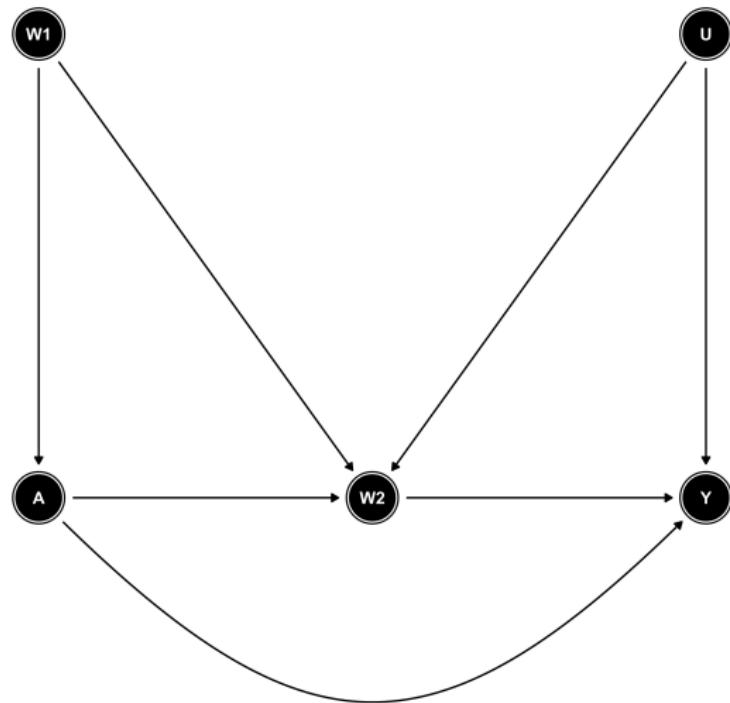
Assumptions:

- (1) is $Y_{am} \perp\!\!\!\perp A|C$
- (2) is $Y_{am} \perp\!\!\!\perp M|C,A$
- (3) is $M_a \perp\!\!\!\perp A|C$
- (4) is $Y_{am} \perp\!\!\!\perp M_a|C$



⁶ Luque-Fernandez M.A., Zoega, H., Valdimarsdottir, U. et al. Eur J Epidemiol (2016) 31: 613. <https://doi.org/10.1007/s10654-016-0139-5>

DAG example: Dagitty and ggdag (LABs)



⁷ Luque Fernandez MA et al. Causal Modeling Triangulation in Epidemiology: From Parametric G-computation to semi-parametric double-robust cross-validated and collaborative TMLE (2019). In press

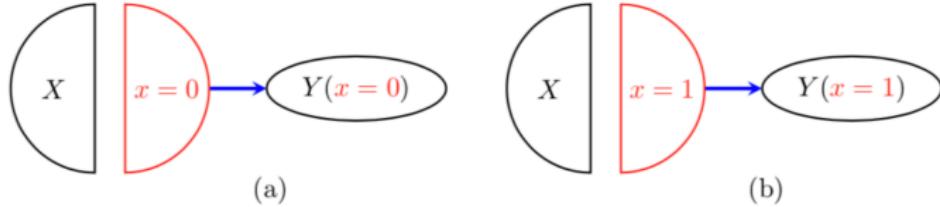


Figure 2: The single world intervention graphs (SWIGs) resulting from splitting node X in the graph in Figure 1(a), and intervening to set a particular value. (a) the SWIG $\mathcal{G}(x=0)$ corresponding to setting X to 0; (b) $\mathcal{G}(x=1)$ given by setting X to 1.

DAGs represent Structural Equation Models

Confounder structure: Simple linear simulation

```
N <- 1000 # sample size  
set.seed(777) # random seed  
W <- rnorm(N) # confounder  
A <- 0.5 * W + rnorm(N) # exposure  
Y <- 0.3 * A + 0.4 * W + rnorm(N) # outcome  
fit1 <- lm(Y ~ A) # crude model  
fit2 <- lm(Y ~ A + W) # adjusted model
```

Collider structure

```
N <- 1000 # sample size  
set.seed(777) # random seed  
A <- rnorm(N) # exposure  
Y <- 0.3 * A + rnorm(N) # outcome  
C <- 1.2 * A + 0.9 * Y + rnorm(N) # collider  
fit3 <- lm(Y ~ A) # crude model  
fit4 <- lm(Y ~ A + C) # adjusted model
```

Collider and confounding effects

		Dependent variable (Y)			
W (confounder)		C (collider)			
	Unadjusted β (SE)	Adjusted β (SE)	Unadjusted β (SE)	Adjusted β (SE)	
A	0.471 (-0.030)	0.289 (-0.032)	A	0.326 (-0.031)	-0.416 (-0.035)
W		0.425 (-0.035)	C		0.491 (-0.018)
Intercept	-0.061 (-0.033)	-0.06 (-0.031)		0.01 (-0.031)	0.035 (-0.023)
AIC	100.42	-31.992		-55.369	-626.824

Note: Lower AIC is better

Luque-Fernandez et al. Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application <https://arxiv.org/abs/1809.07111>

Display Linear Fit: models (fit2) and (fit4)

Figure 2A

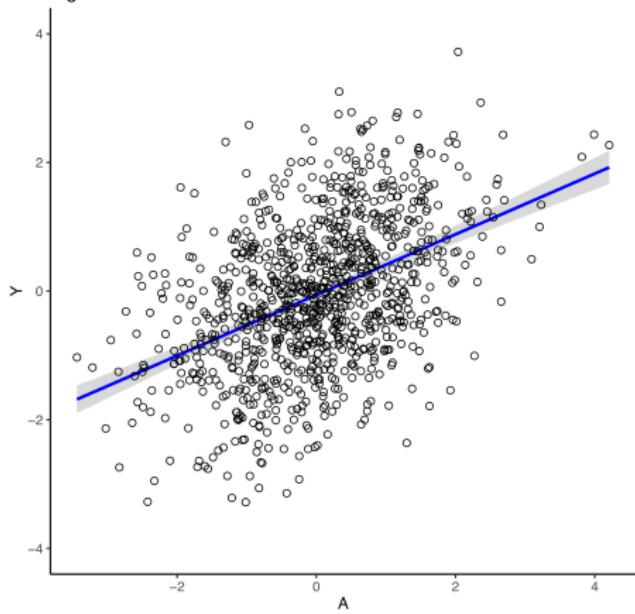
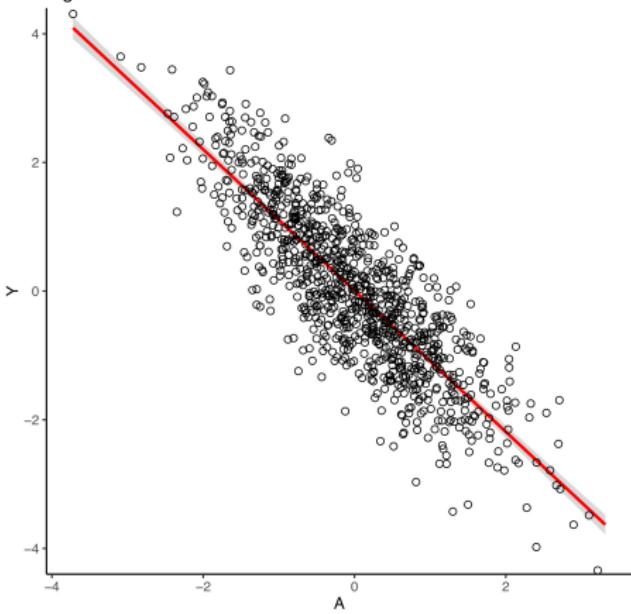


Figure 2B



Collider Effect

Luque-Fernandez et al.(2018) Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application <https://arxiv.org/abs/1809.07111>

Shiny web application

Colliders in Epidemiology: an educational interactive web application

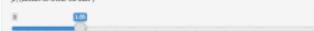
Motivation Data generation Collider Visualization Article Credits & Acknowledgment

Share it on Twitter

Effect of dietary sodium intake on systolic blood pressure for different models' specifications.

Move the slider to change the magnitude of the true causal effect of sodium in SBP

Causal Model:
 β_1 (Effect of SOD on SBP)



Collider Model:
 $\text{PRO} = \alpha_1 + \alpha_2 \text{SOD} + \alpha_3 \text{SBP}$

Move the sliders to change the magnitude of the effect of sodium and SBP in proteinuria

α_1 (Effect of SOD on PRO)



α_2 (Effect of SBP on PRO)



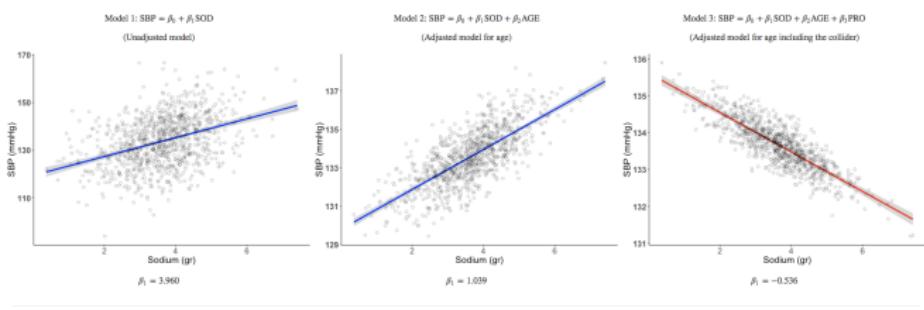
Legend:
AGE = Age (years)
SOD = 24-hour dietary sodium intake (g)
PRO = 24-hour excretion of urinary protein (proteinuria) (mg)
SBP = Systolic blood pressure (mmHg)

Select the model(s) to visualize the effect of SOD in SBP:

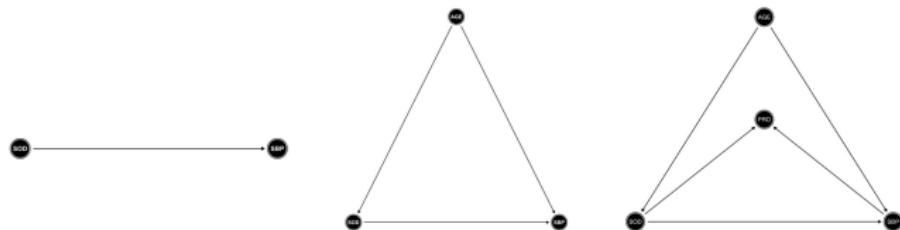
SBP = $\beta_0 + \beta_1 \text{SOD}$

SBP = $\beta_0 + \beta_1 \text{SOD} + \beta_2 \text{AGE}$

SBP = $\beta_0 + \beta_1 \text{SOD} + \beta_2 \text{AGE} + \beta_3 \text{PRO}$

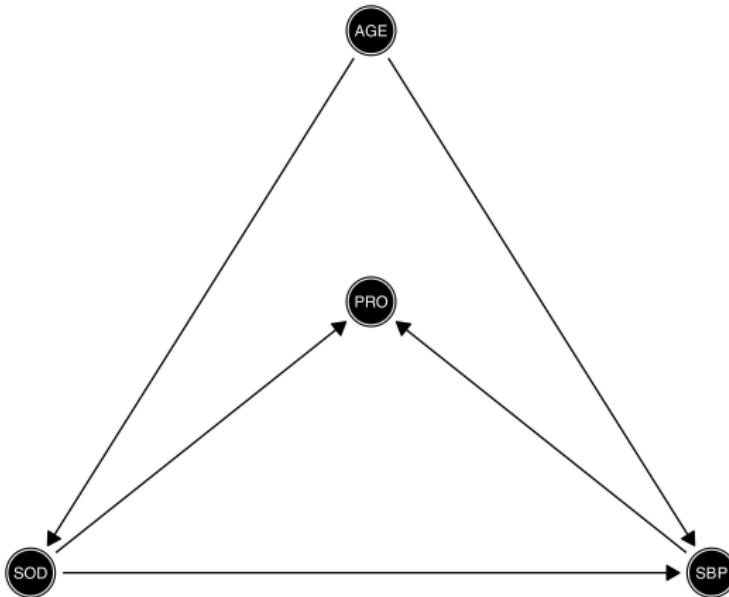


Assumed DAG under respective model



Colliders in Epidemiology: an educational interactive Shiny web application

Example Simulation using a DAG (LAB4)



Directed acyclic graph depicting the structural causal relationship of the exposure and outcome, confounding and collider effects.
Exposure: 24-hour sodium dietary intake in gr (SOD), outcome: systolic blood pressure in mmHg (SBP), confounder: age in years (AGE), collider: 24-hour urinary protein excretion, proteinuria (PRO).

Luque-Fernandez et al. (2018) Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application <https://arxiv.org/abs/1809.07111>



Seeting Monte Carlo simulations

Data Generation

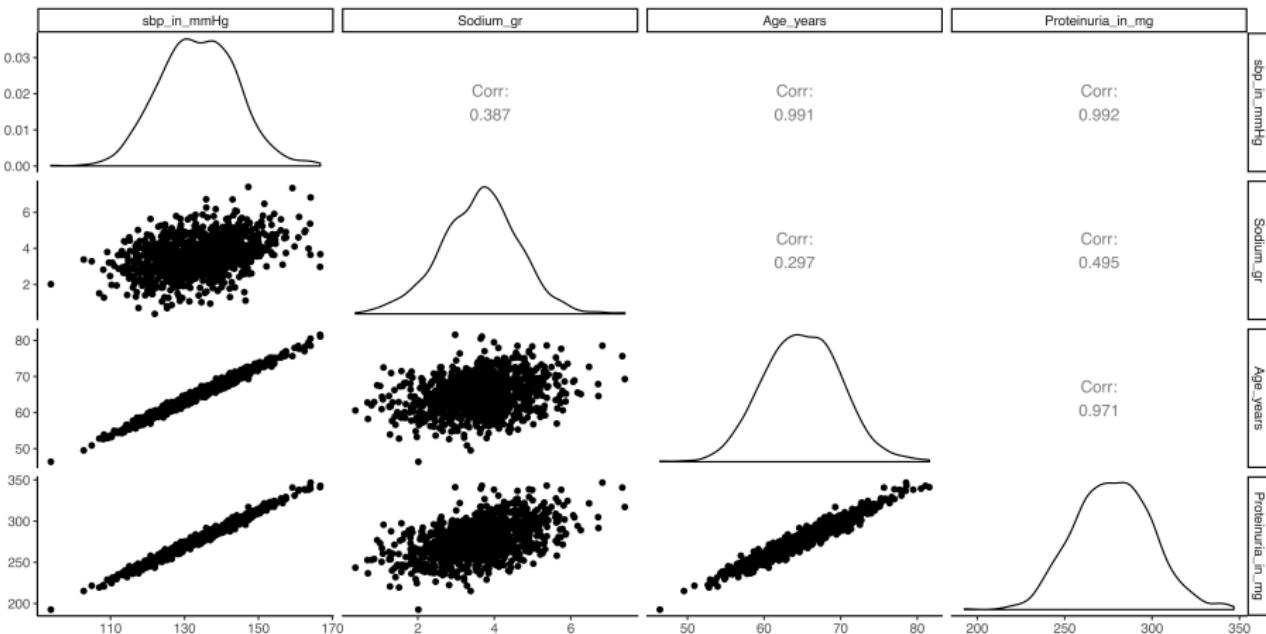
```
generateData <- function(n, seed) {  
  set.seed(seed)  
  Age_years <- rnorm(n, 65, 5)  
  Sodium_gr <- Age_years / 18 + rnorm(n)  
  sbp_in_mmHg <- 1.05 * Sodium_gr + 2.00 * Age_years + rnorm(n)  
  hypertension <- ifelse(sbp_in_mmHg>140,1,0)  
  Proteinuria_in_mg <- 2.00*sbp_in_mmHg + 2.80*Sodium_gr + rnorm(n)  
  data.frame(sbp_in_mmHg, hypertension, Sodium_gr, Age_years,  
             Proteinuria_in_mg)  
}  
ObsData <- generateData(n = 1000, seed = 777)
```

Monte Carlo simulations

MC simulations

```
R<-1000
true <- rep(NA, R)
collider <- rep(NA,R)
se <- rep(NA,R)
set.seed(050472)
for(r in 1:R) {
  if (r%%10 == 0) cat(paste("This is simulation run number", r, "\n"))
  ObsData <- generateData(n=10000)
  # True effect
  true[r] <- summary(lm(sbp_in_mmHg ~ Sodium_gr + Age_years, data = ObsData))$coef[2,1]
  # Collider effect
  collider[r] <- summary(lm(sbp_in_mmHg ~ Sodium_gr + Age_years + Proteinuria_in_mg,
    data = ObsData))$coef[2,1]
  se[r] <- summary(lm(sbp_in_mmHg ~ Sodium_gr + Age_years + Proteinuria_in_mg, data = ObsData))$coef
}
# Estimate of sodium true effect
mean(true)
# Estimate of sodium biased effect in the model including the collider
mean(collider)
# simulated standard error/confidence interval of outcome regression
lci <- (mean(collider) - 1.96*mean(se)); mean(lci)
uci <- (mean(collider) + 1.96*mean(se)); mean(uci)
# Bias
Bias <- (true - abs(collider));mean(Bias)
# % Bias
relBias <- ((true - abs(collider)) / true); mean(relBias) * 100
# Plot bias
plot(relBias)
```

One sample MC simulations



Visualization of the multivariate structure of the data generation, $n = 1,000$.

Luque-Fernandez et al.(2018) Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application <https://arxiv.org/abs/1809.07111>



Models specifications

Unadjusted model

$$\text{SBP in mmHg} = \beta_0 + \beta_1 \times \text{Sodium in gr} + \varepsilon$$

Adjusted model (confounder)

$$\text{SBP in mmHg} = \beta_0 + \beta_1 \times \text{Sodium in gr} + \beta_2 \times \text{Age in years} + \varepsilon$$

Adjusted model (confounder and collider)

$$\text{SBP} = \beta_0 + \beta_1 \times \text{Sodium} + \beta_2 \times \text{Age} + \beta_3 \times \text{Proteinuria} + \varepsilon$$

Models fit visualization

Figure 4A

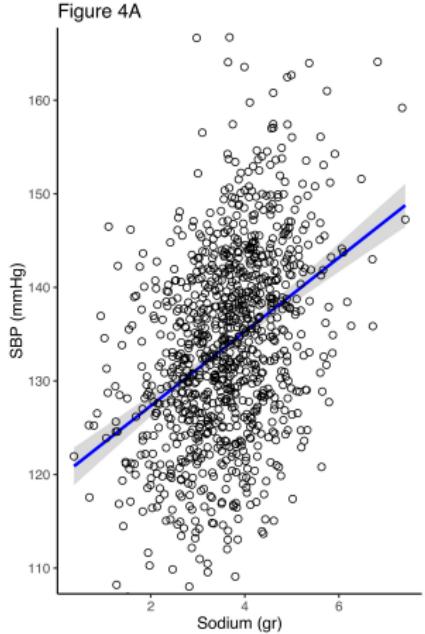


Figure 4B

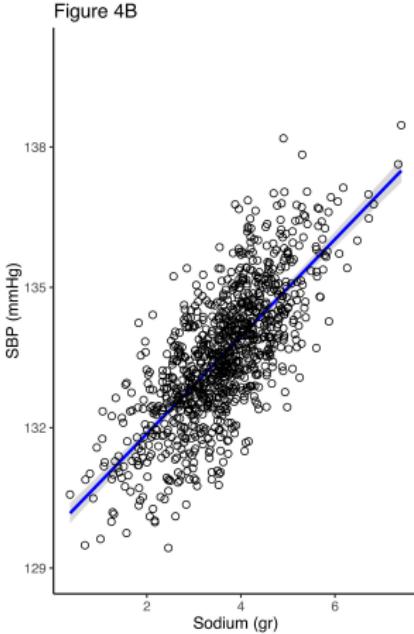
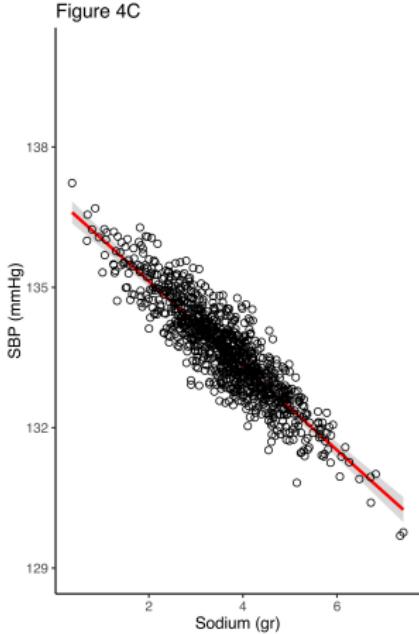


Figure 4C



Luque-Fernandez et al.(2018) Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application <https://arxiv.org/abs/1809.07111>

Collider and confounding effects

	Dependent variable: SBP in mmHg		
	Univariate	Bivariate	Multivariate
	(SE)	(SE)	(SE)
True effect of Sodium in gr: 1.05			
Sodium in gr	3.960 (0.298)	1.039 (0.032)	-0.902 (0.036)
Age in years		2.004 (0.007)	0.416 (0.027)
Proteinuria in mg			0.396 (0.007)
Intercept	119.420 (1.122)	-0.311 (0.407)	-0.091 (0.192)
AIC	7363.45	2807.89	1302.66

Note: Lower AIC is better

pdf

Luque-Fernandez et al. Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application <https://arxiv.org/abs/1809.07111>



Colliders Tutorial

Collider Shiny App

<http://watzilei.com/shiny/collider/>

GitHub Open source Collider files

<https://github.com/migariane/ColliderApp>

DAGs software: LABs 3 and 4

DAGITTY

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

GGDAG

<https://ggdag.netlify.com/articles/intro-to-dags.html>

Thank you!



INSTITUTO DE INVESTIGACIÓN BIOSANITARIA



"Una manera de hacer Europa"

Miguel Angel Luque-Fernandez

miguel.luque.easp@juntadeandalucia.es

@watzilei

Carlos III Institute of Health, Grant/Award Number: **CP17/00206**