ibs.GRANADA
INSTITUTO DE INVESTIGACIÓN BIOSANITARIA

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD
Instituto de Salud Carlos III
UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional
"Una manera de hacer Europa"

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

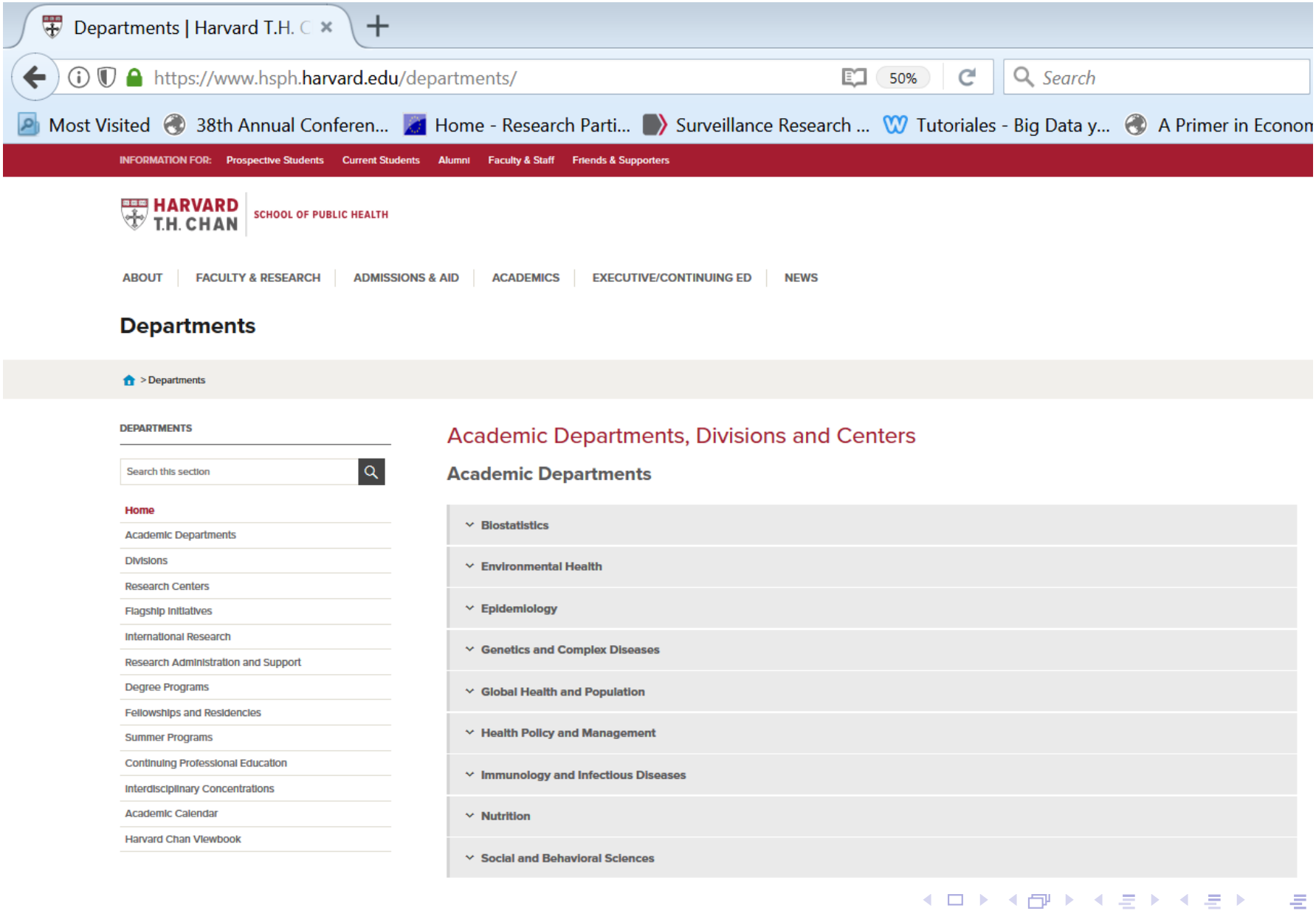
Miguel Angel Luque Fernandez

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



Epidemiology as scientific method: areas of concentration

Department of Epidemiology

https://www.hsph.harvard.edu/epidemiology/

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

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 frequ
and determinants of disease in humans, a fundamental science of public health. In addition to pursuing groundbre
research initiatives, we educate and prepare future medical leaders and practitioners as part of our mission to ignit
changes in the quality of health across the world.

Innovative Educational Programs: Onsite and Online

Data Science Initiative

o-directors of newly launch

https://news.harvard.edu/gazette/story/2017/03/co-directors-of-newly-launched

Visited | 38th Annual Conferen... | Home - Research Parti... | Surveillance Research ... | Tutoriales - Big Data y... | A Primer in Economet... | GESTIÓN DE PROYECT... | DeepL Translato

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

"Most important is to create opportunities for scientists here to interact in new ways," said Francesca Dominici (1), co-director of the Harvard Data Science Initiative. "We will succeed if we can get people who are working on data-science-related topics all across the University to get to know each other better." Dominici, senior associate dean for research at the Harvard Chan School, will co-direct the initiative with David C. Parkes (2), George F. Colony Professor and area dean for computer science at SEAS.

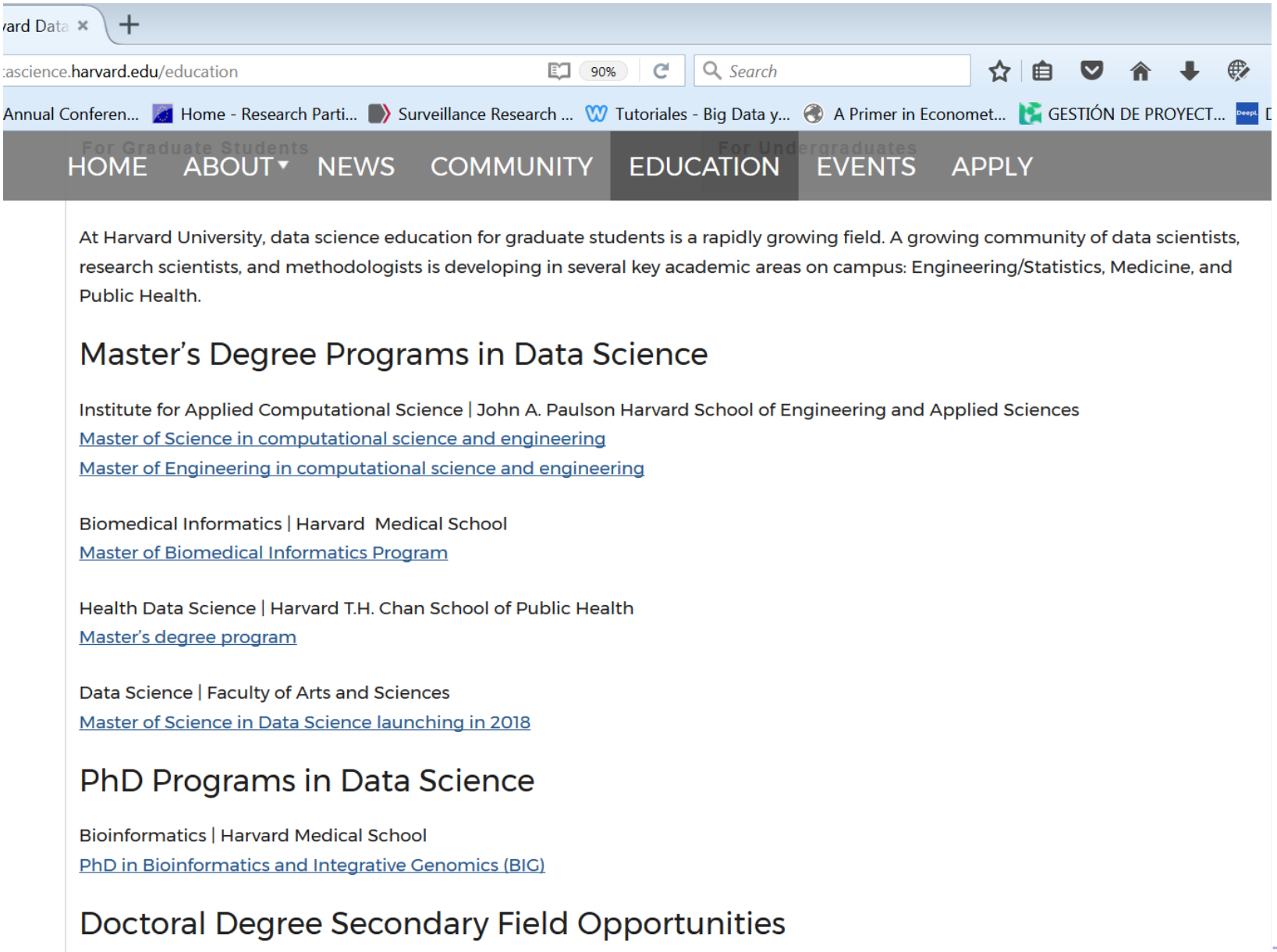
Harvard University just announced the launch of its Data Science Initiative, a program to harness the vast expertise and innovations that are occurring in disciplines as diverse as medicine, law, policy, and computer science.

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

POPULAR

Data Science Programmes



ard Data

ascience.harvard.edu/education 90% Search

Annual Conferen... Home - Research Parti... Surveillance Research ... Tutoriales - Big Data y... A Primer in Economet... GESTIÓN DE PROYECT...

For Graduate Students For Undergraduates

HOME ABOUT NEWS COMMUNITY EDUCATION EVENTS APPLY

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

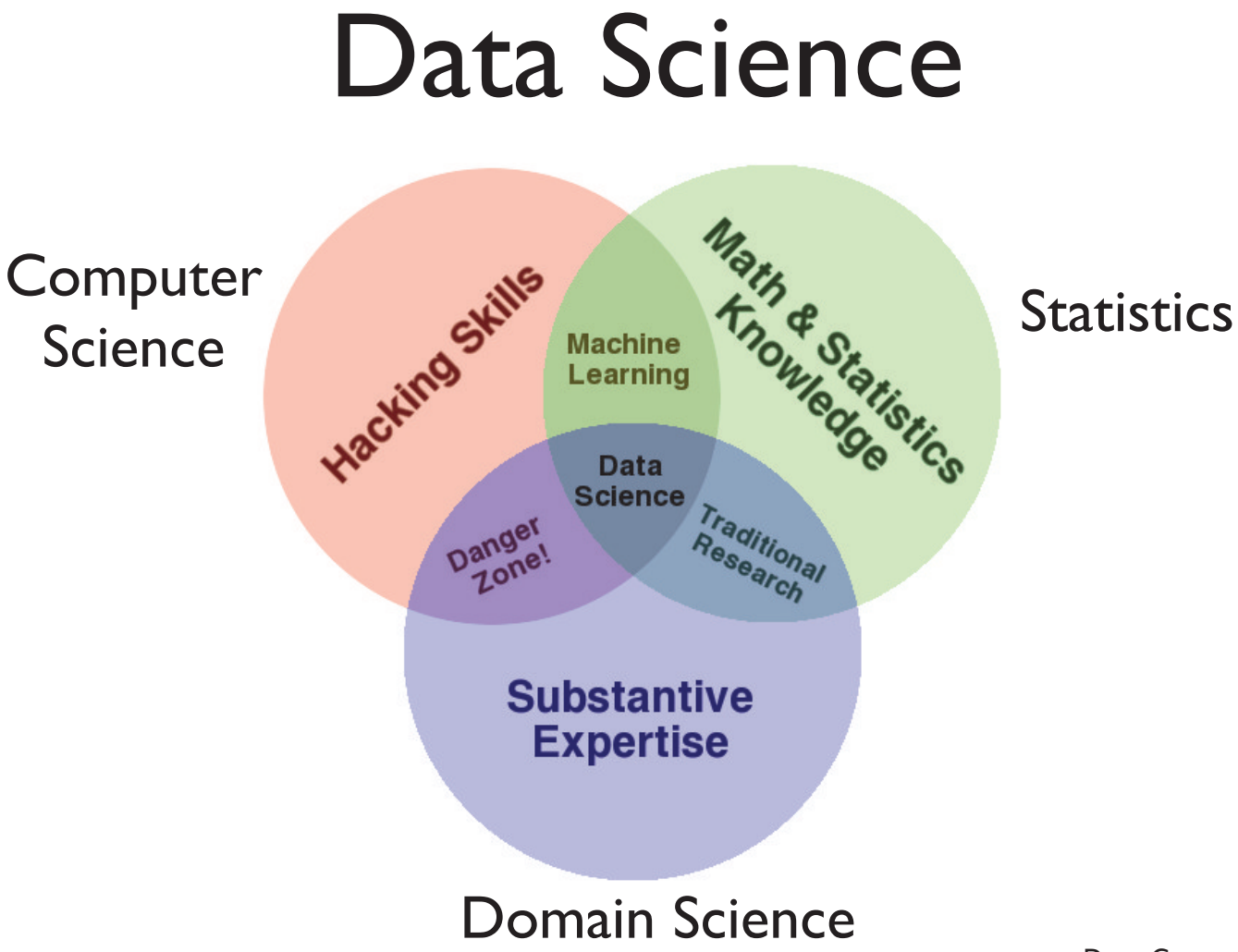
Data Science | Faculty of Arts and Sciences
Master of Science in Data Science launching in 2018

PhD Programs in Data Science

Bioinformatics | Harvard Medical School
PhD in Bioinformatics and Integrative Genomics (BIG)

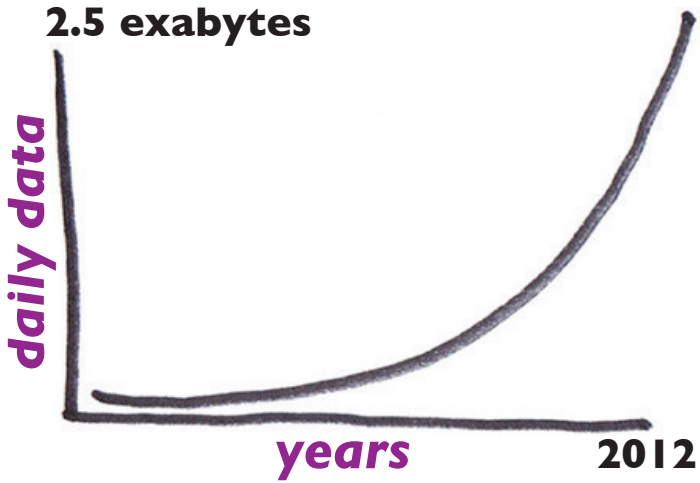
Doctoral Degree Secondary Field Opportunities

Data Science Domains

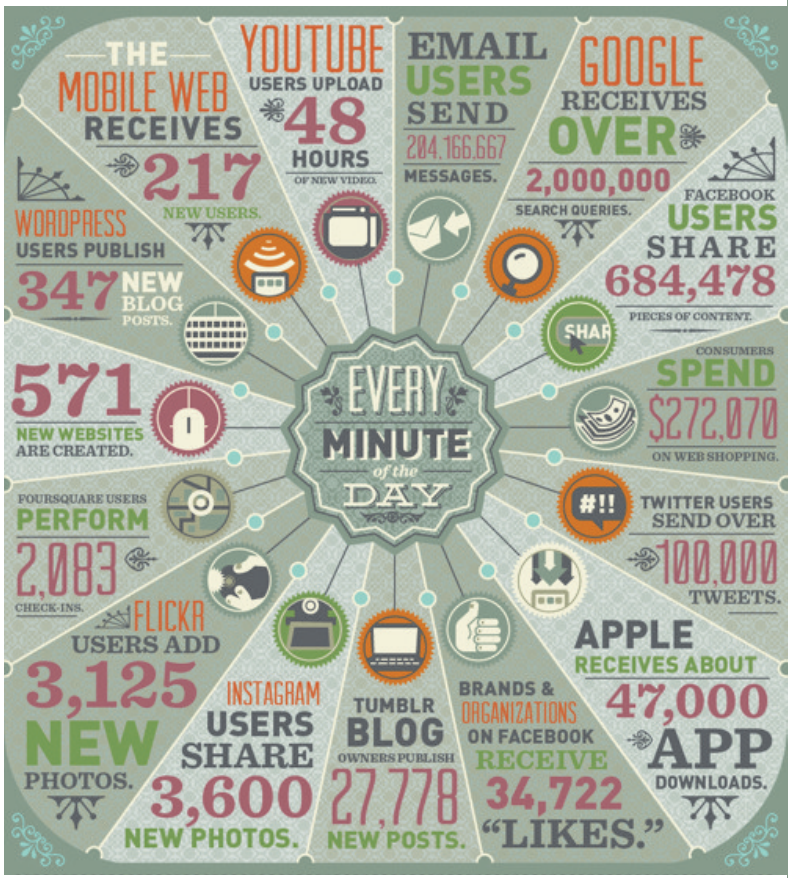


Drew Conway

Big Data

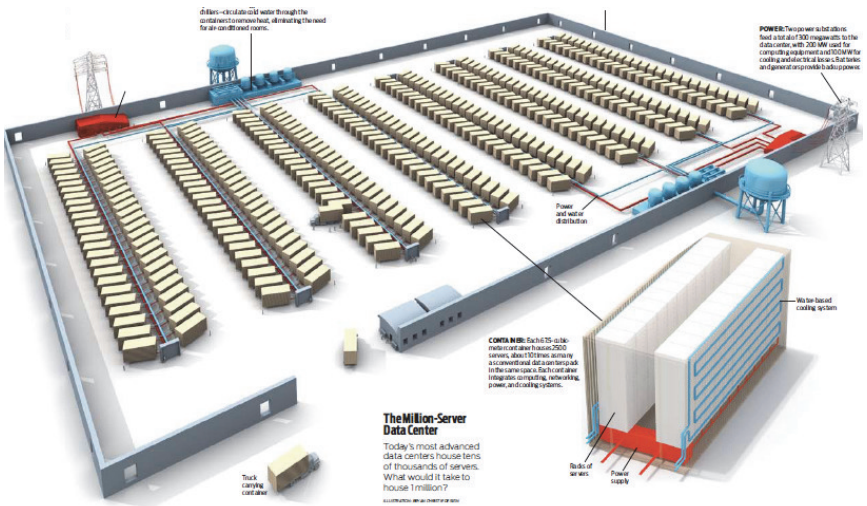


[IBMbigdata]



[Domo]

Commodity Computing



Michael Franklin, UC Berkeley

Data Science and data sources: Variety

Smarter Devices

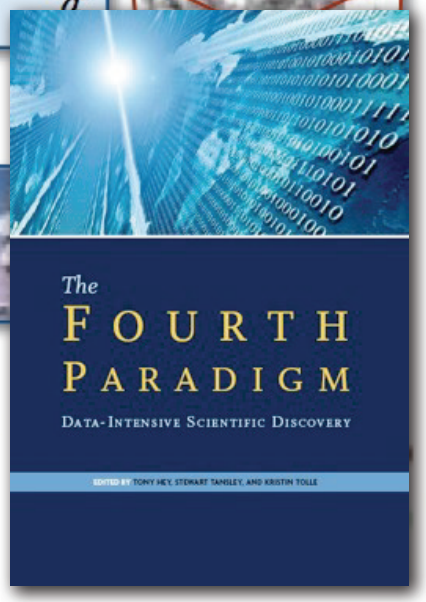
Michael Franklin, UC Berkeley

New Science Paradigm

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

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



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"

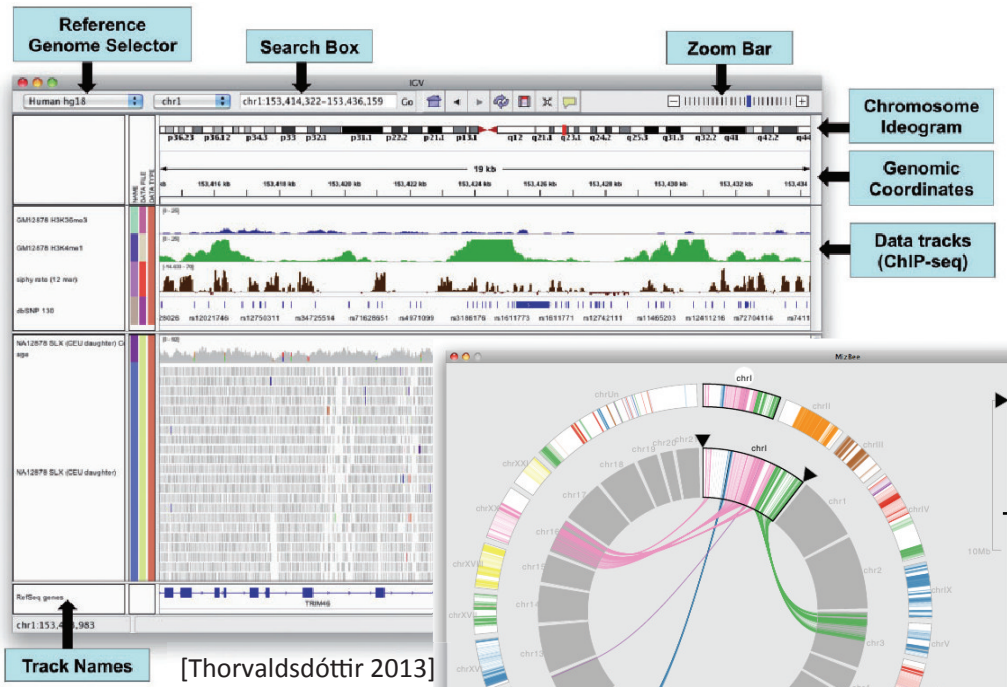
McKinsey Global Institute

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

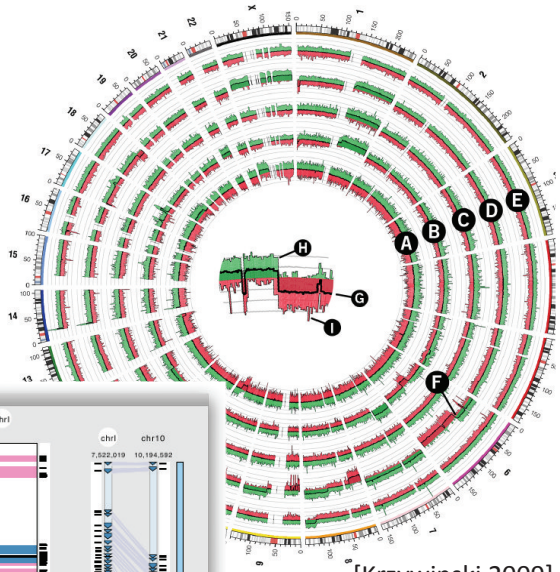
Hal Varian, Prof. Emeritus UC Berkeley
Chief Economist, Google

So, how about Epidemiology?

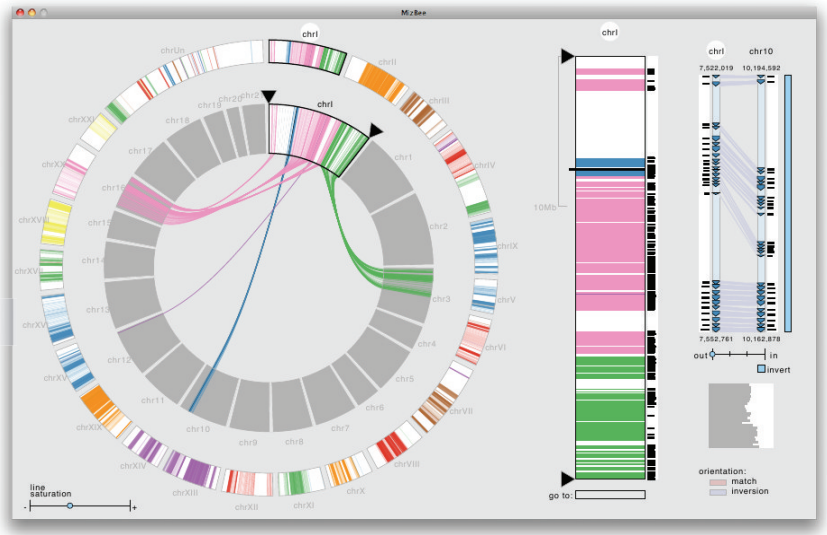
Genome Visualization



[Thorvaldsdóttir 2013]



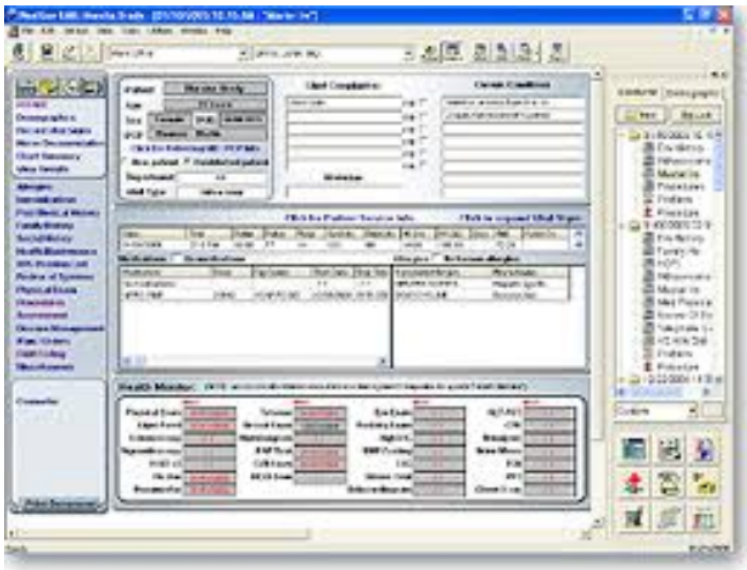
[Krzywinski 2009]



[Meyer 2009]

So, how about Epidemiology?

electronic charting | health record | digital | patient | nursing | doctor | c



So, how about Epidemiology?

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** (...)."

ELSEVIER

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 Hos 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.

Keywords: Utilization databases; Claims data; Therapeuties; 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].

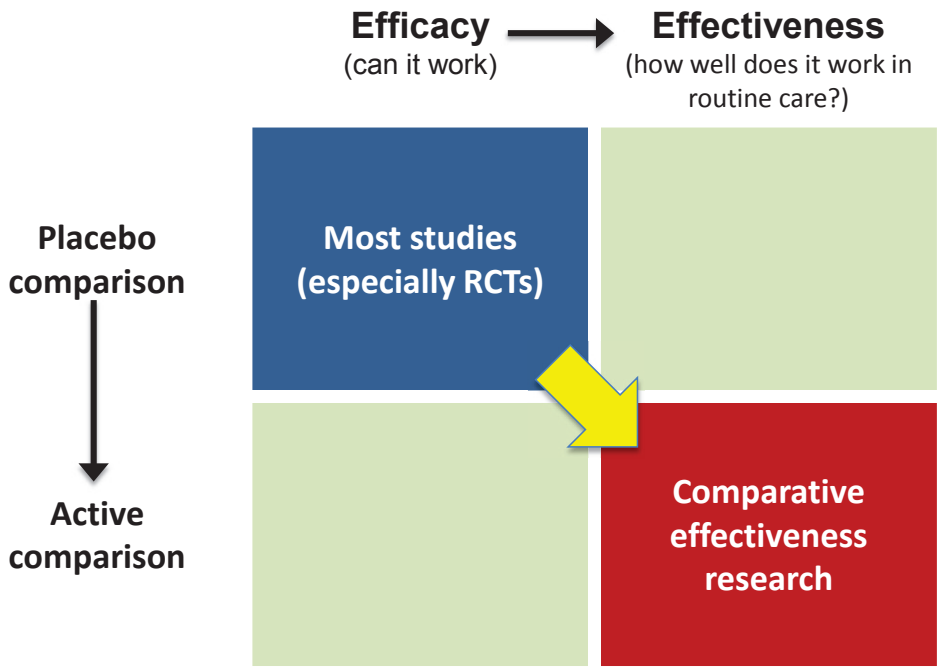
Although pharmacoepidemiology makes use of all epidemiologic study designs and data sources, in recent years there has been enormous growth in the use of large health care databases [5]. These are made up of the automated electronic recording of filled prescriptions, professional services, and hospitalizations; 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

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.

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://wwwold.academyhealth.org/files/FileDownloads/AH_Monograph_09FINAL7.pdf

CER is different

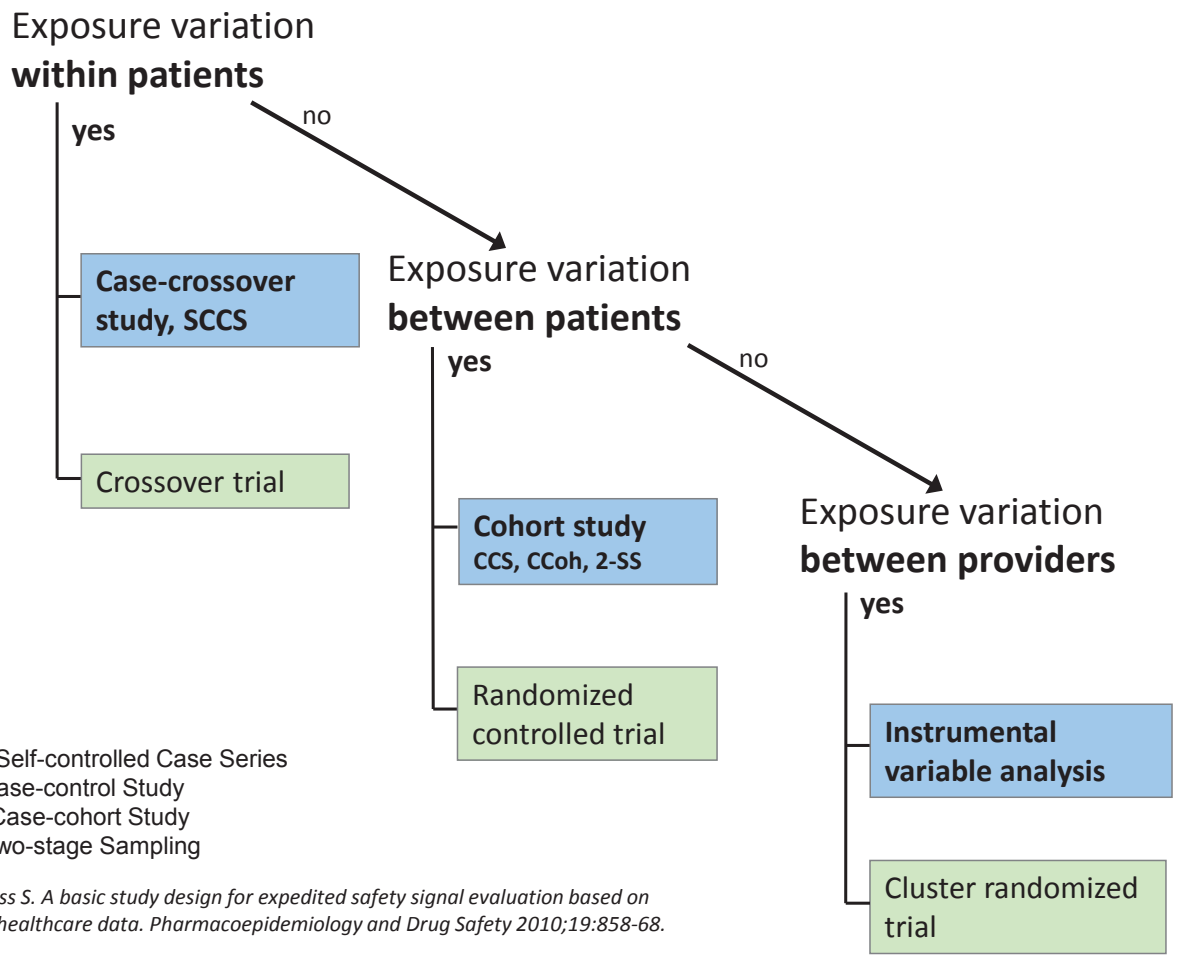
What CER seeks to do

	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

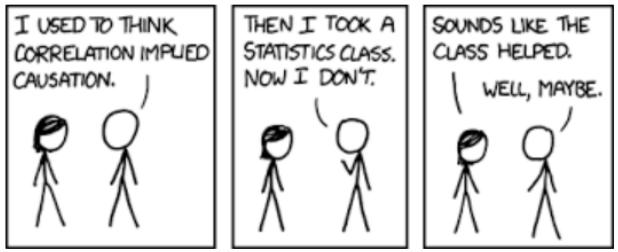
Design choice: Source of exposure variation



SCCS: Self-controlled Case Series
CCS: Case-control Study
CCoh: Case-cohort Study
2-SS: Two-stage Sampling

Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiology and Drug Safety 2010;19:858-68.

CER is about Causal Inference



CAUSAL INFERENCE IN STATISTICS

A Primer

Judea Pearl
Madelyn Glymour
Nicholas P. Jewell

WILEY

Please do not distribute without permission

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 **Y**, or the "outcome" or "response" or variable.
- The **independent variable** measures that variation which we think explains variation in the dependent variable.
  - Also called **X**, or the "treatment" or "study" variable.

Estimating causal effects from observational data

Why is causal inference hard?
- Simpson's paradox

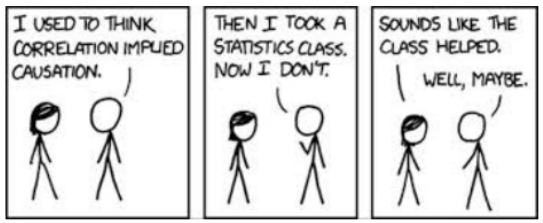
The language of graphical models
- Backdoor criterion
- Frontdoor criterion

Common approaches for causal inference
- Conditioning
- Mechanism-based
- Natural Experiments

Example: Estimating causal impact of recommender systems

Model for Causal Inf

- For causal questions, we wi
  happen if a policy-maker ch
  - Potential outcomes notation:
    - Y(w) is the outcome unit i wo
    - For binary treatment, treatme
  - Administer a drug, change mi
  - Function of interest: mapping
  - Holland: Fundamental Proble
    - We do not see the same units
- Units of study typically have
  - These would not change with
  - E.g. we don't contemplate mo
    we change minimum wage po

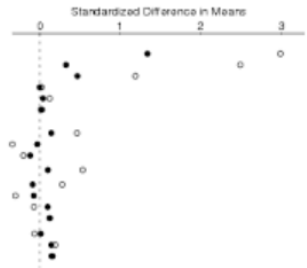


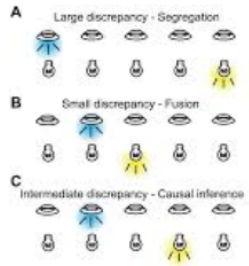
A Research Agenda on Causal Inference

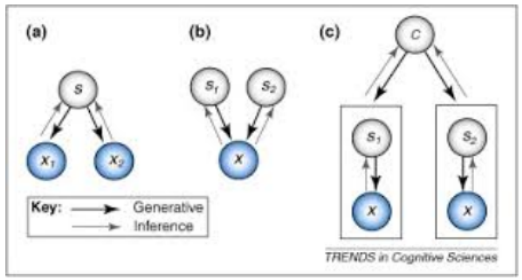
Problems
- Many problems in social sciences entail a combination of prediction and causal inference
- Existing ML approaches to estimation, model selection and robustness do not directly apply to the problem of estimating causal parameters
- Inference more challenging for some ML methods

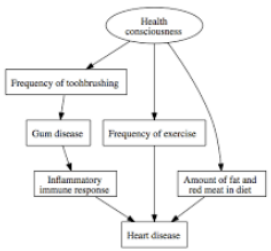
Proposals
- Formally model the distinction between causal and predictive parts of the model and treat them differently for both estimation and inference
  - Abadie, Athey, Imbens and Wooldridge (2014, under review; also work in progress)
- Develop new estimation methods that combine ML approaches for prediction component of models with causal approaches
  - Athey-Imbens (2015, work in progress)
- Develop new approaches to cross-validation optimized for causal inference and optimal policy estimation
  - Athey-Imbens (2015, work in progress)
- Develop robustness measures for causal parameters inspired by ML
  - Athey-Imbens (AER P&P 2015, work in progress)
- Develop methods for causal inference for network analysis drawing on CS tools for networks
  - Athey-Eckles-Imbens (2015)
- Large scale structural models with latent variables
  - Athey-Nekipelov (2012, 2015); Athey, Biel, Hofman

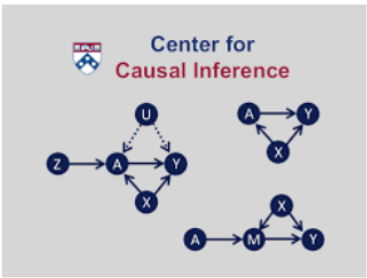
Predictor	
Child	
	negative birth weight
	weeks preterm
	days in hospital
	male
	first born
	age
Mother	
	black
	hispanic
	white
	unmarried at birth
	less than high school
	high school graduate
	some college
	college graduate
	worked during pregnancy
	had no prenatal care
	age at birth





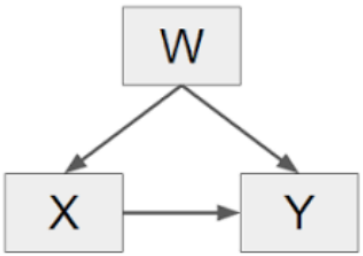






Casual Inference Diagrams

Parametric Approaches

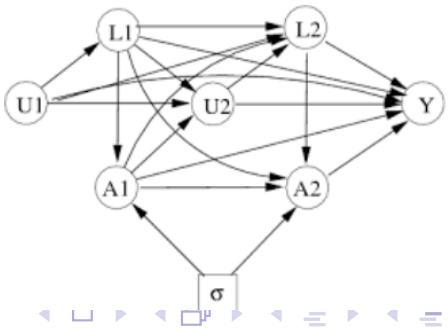


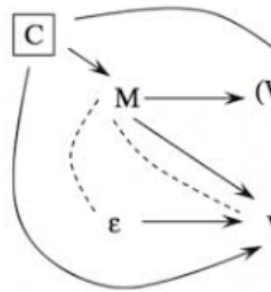
Causal Inference

- 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?

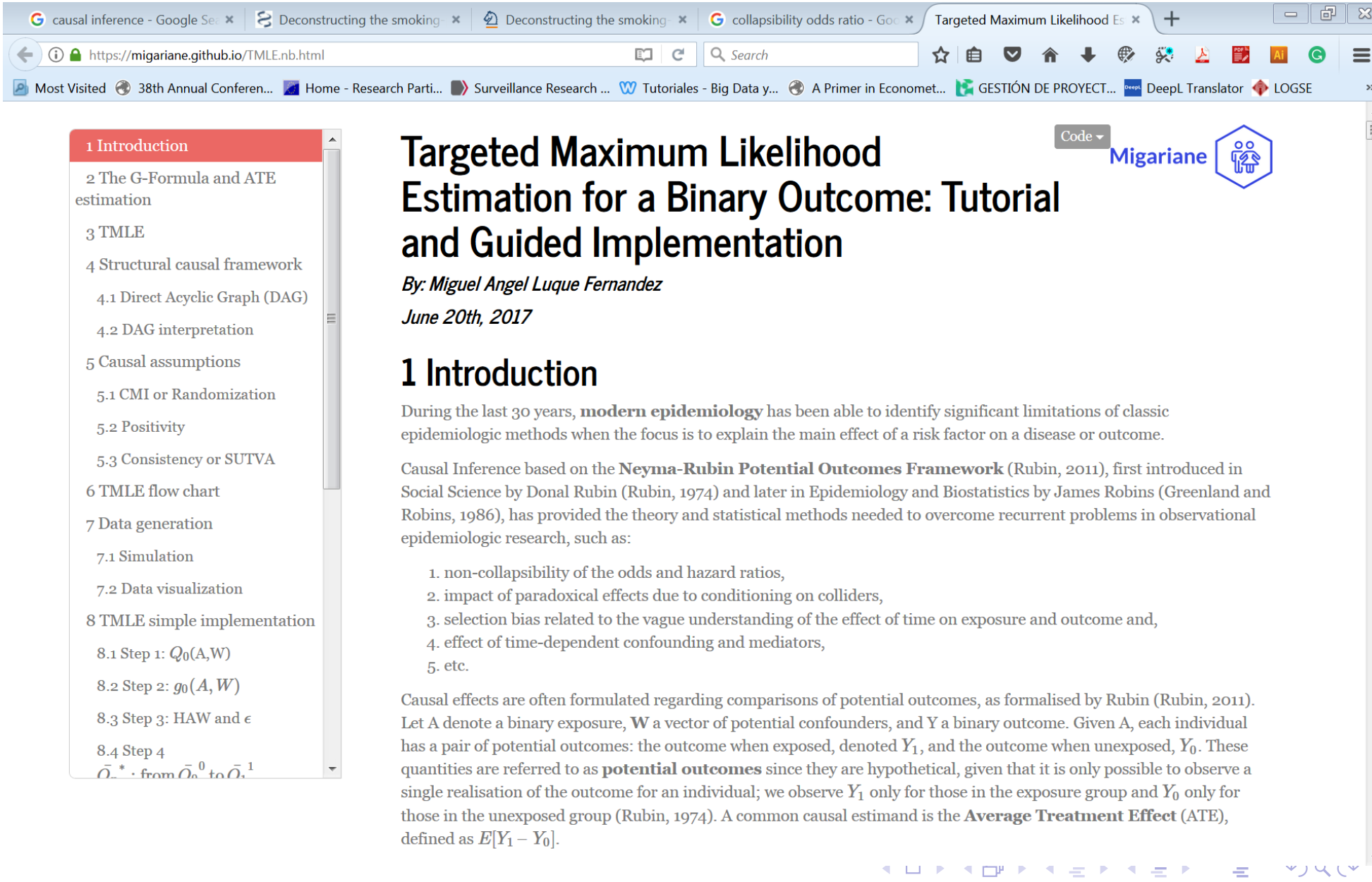
Headlines

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

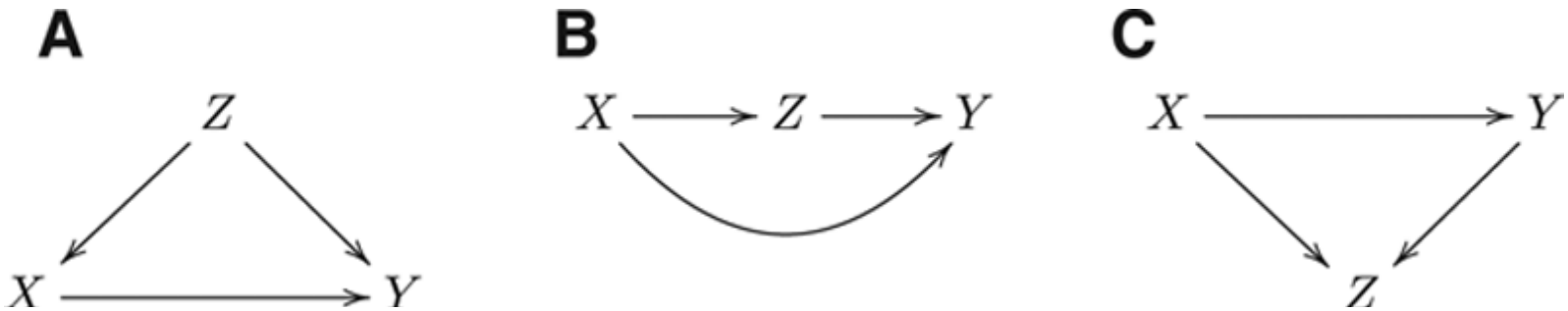




CER is about Causal Inference



CER is about Causal Inference

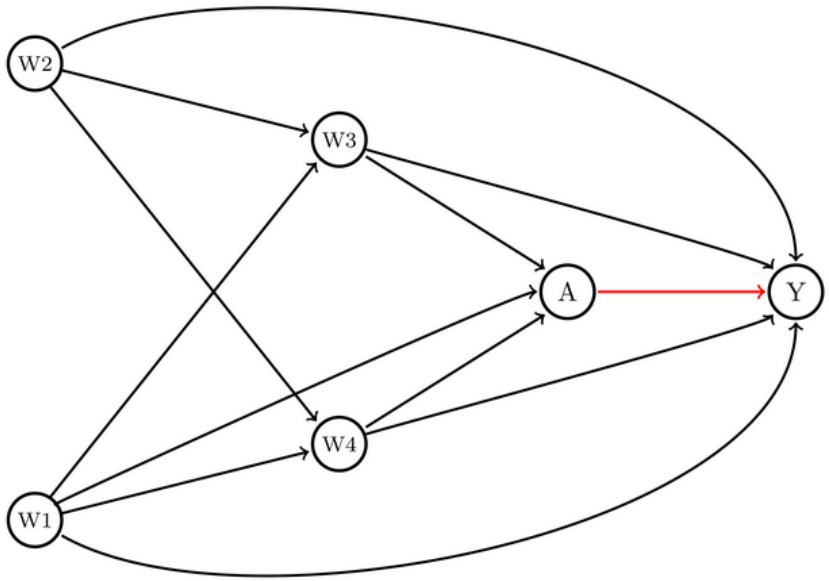


CER is about Causal Inference

Direct Acyclic Graph (DAG)

Under conditional exchangeability: $Y(0), Y(1) \perp A|W$

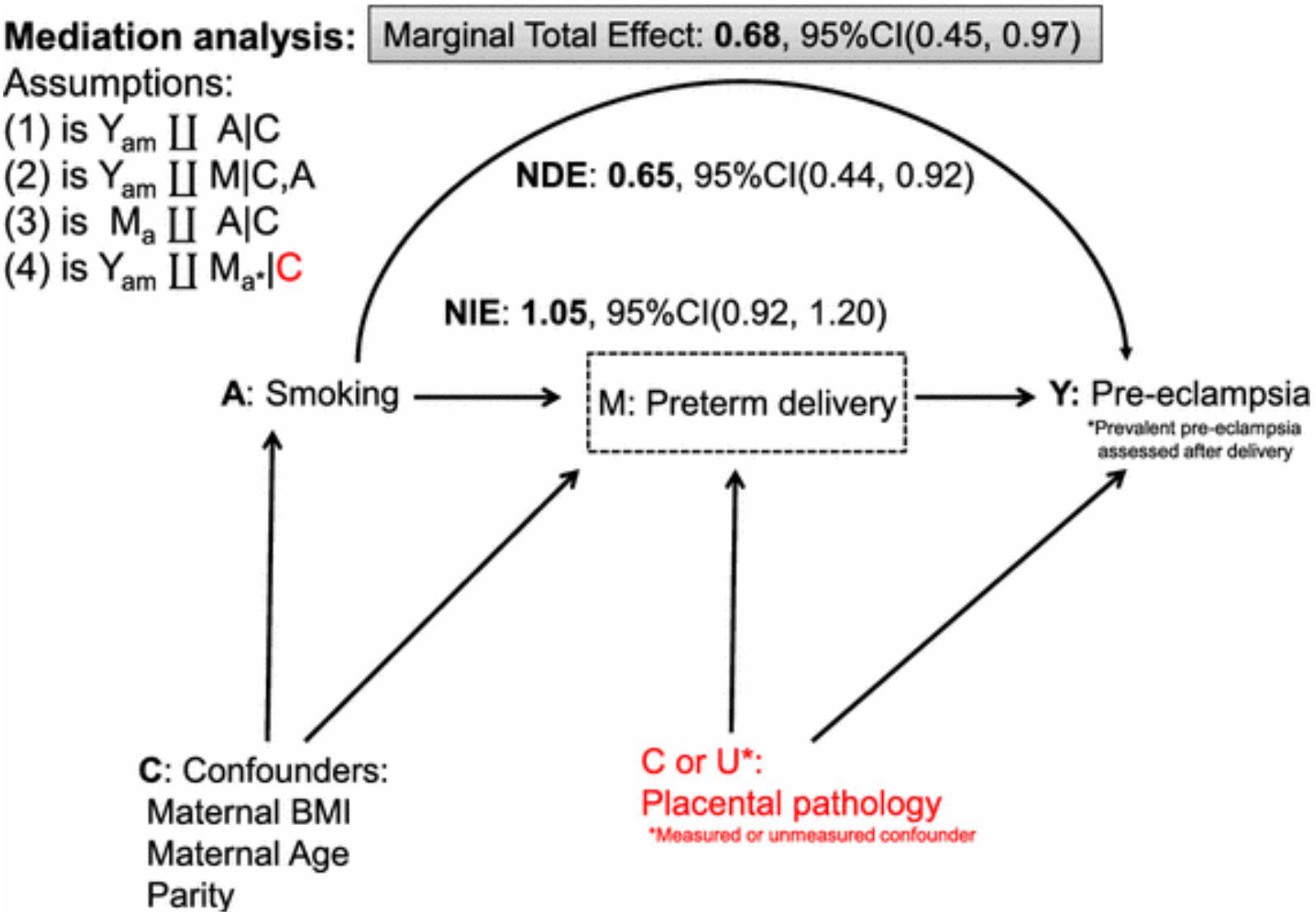
$$\text{ATE} = E[E(Y|A=1;W) - E(Y|A=0;W)]$$



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

[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



[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 Dahlqwist, 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

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

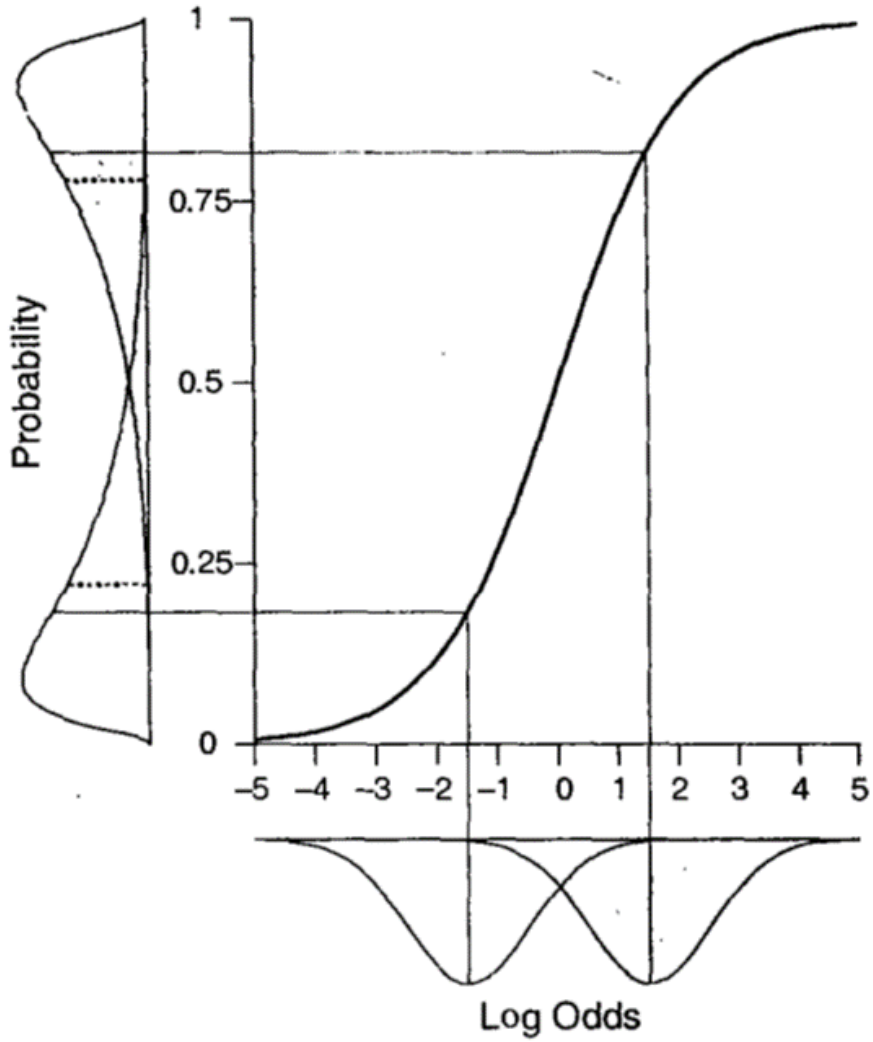
that the conditional odds ratio is constant across levels of Z (e.g., in logistic regression with main effects only).

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

$$\frac{\Pr(Y=1\mid X=1)\Pr(Y=0\mid X=0)}{\Pr(Y=0\mid X=1)\Pr(Y=1\mid X=0)}.$$

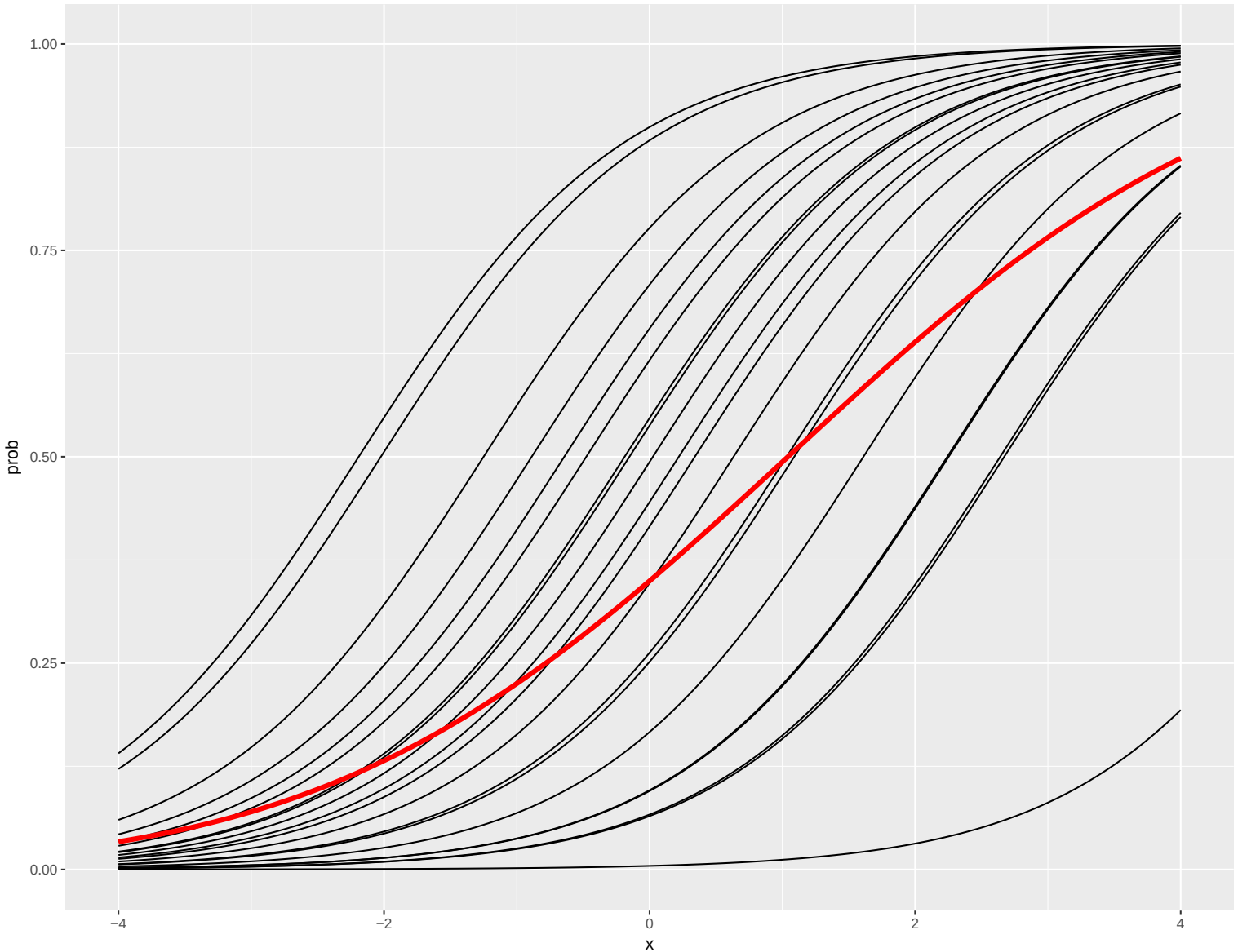
One explanation for a discrepancy between the conditional and marginal odds ratio could be that Z is a confounder (Fig. 1A); this would typically be the argument for adjusting for Z in the first place. Other explanations could be that Z is a mediator (Fig. 1B) or a collider (Fig. 1C). All these explanations 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] Neuhaus and Jewell[2] showed that the mar-

CER is about Causal Inference



Fitzmaurice (2011). Applied Longitudinal Analysis

CER is about Causal Inference



[4] MA LUQUE-FERNANDEZ EPM304: Non-linear Random Effects Models

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.

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.

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**"

Annals of Internal Medicine

The Spectrum of Subclinical Prima 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-

(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-

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.

Annals of Internal Medicine

The Spectrum of Subclinical Prima Hypertension

A Cohort Study

Jenifer M. Brown, MD; Cassianne Robinson-Cohen, PhD; Migu Matthew A. Allison, MD, MPH; Rene Baudrand, MD; Joachim H 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-

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

Yoosoo Chang, MD, PhD; Seungho Ryu, MD, PhD; Yuni Choi, BS; Yiyi Zhang, PhD; Juhee Cho, PhD; Min-Jung Kwon, MD, PhD; Young Youl Hyun, MD, PhD; Kyu-Beck Lee, MD, PhD; Hyang Kim, MD, PhD; Hyun-Suk Jung, MD; Kyung Eun Yun, MD, PhD; Jiin Ahn, MSPH; Sanjay Rampal, MD, PhD; Di Zhao, PhD; Byung-Seong Suh, MD, PhD; Eun Cheol Chung, MD, PhD; Hocheol Shin, MD, PhD; Roberto Pastor-Barriuso, PhD; and Eliseo Guallar, 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: Metabolic health was defined as a homeostasis model assessment of insulin resistance less than 2.5 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/m^2, 18.5 to 22.9 kg/m^2, 23 to 24.9 kg/m^2, and 25 kg/m^2 or greater, respectively. The outcome was incident CKD, defined as an estimated glomerular filtration rate less than 60 mL/min/1.73 m^2.

Results: During 369 088 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 normal-weight participants were −4.0 (95% CI, −7.8 to −0.3), 3.5 (CI, 0.9 to 6.1), and 6.7 (CI, 3.0 to 10.4) cases per 1000 persons, respectively. These associations were consistently seen in all clinically relevant subgroups.

Limitation: Chronic kidney disease was identified by a single measurement at each visit.

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 obese phenotype, regardless of metabolic abnormalities, can adversely affect renal function.

Primary Funding Source: None.

Ann Intern Med. 2016;164:305-312. doi:10.7326/M15-1323 **www.annals.org**
For author affiliations, see end of text.
This article was published at www.annals.org on 9 February 2016.

Chronic kidney disease (CKD) is a major clinical and public health problem (1). It is a precursor for end-stage renal disease and a strong risk factor for cardiovascular morbidity 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, hyperglycemia, dyslipidemia, and other metabolic abnormalities–is a major risk factor for CKD (4).

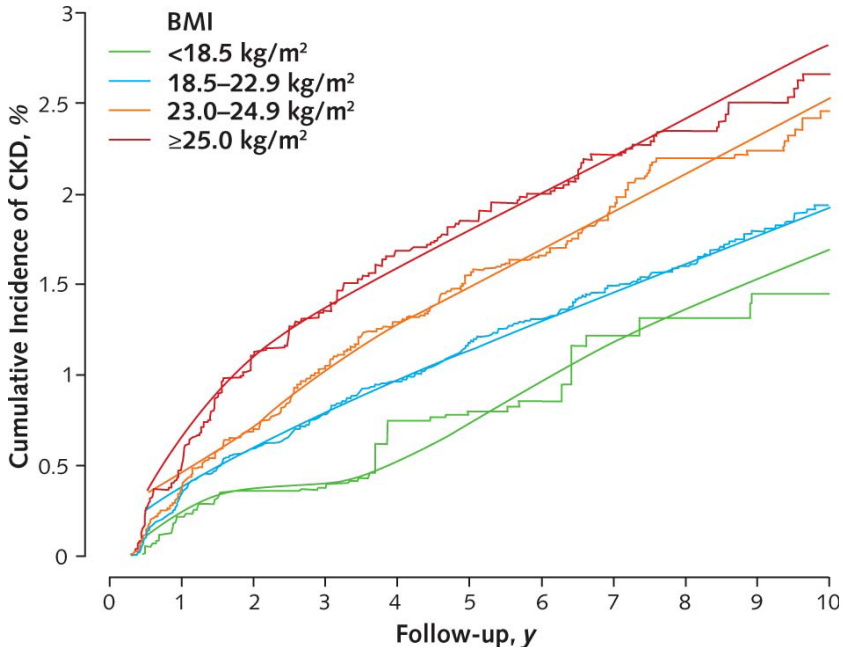
Although the role of obesity-induced metabolic abnormalities in CKD development is well-established, *metabolically healthy obese* (MHO) persons, seem to have a favorable profile with no metabolic abnormalities (5, 6). The association between MHO and CKD, however, is largely unknown. The only study available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

METHODS

Study Population

The Kangbuk Samsung Health Study is a cohort study of South 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 companies and local governmental 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 registered 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 1 other screening examination before 31 December 2013 (that is, they all had a baseline visit and ≥1 follow-up visit [n = 175 859]) (**Figure 1**). We excluded persons who had metabolic abnormalities (5, 9, 10) or evidence of kidney disease at baseline (n = 108 263). We excluded those with fasting glucose levels of 100 mg/dL or greater or who used glucose-lowering agents; blood pressure (BP) of



Background Causal Modelling: Potential Outcomes

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**.

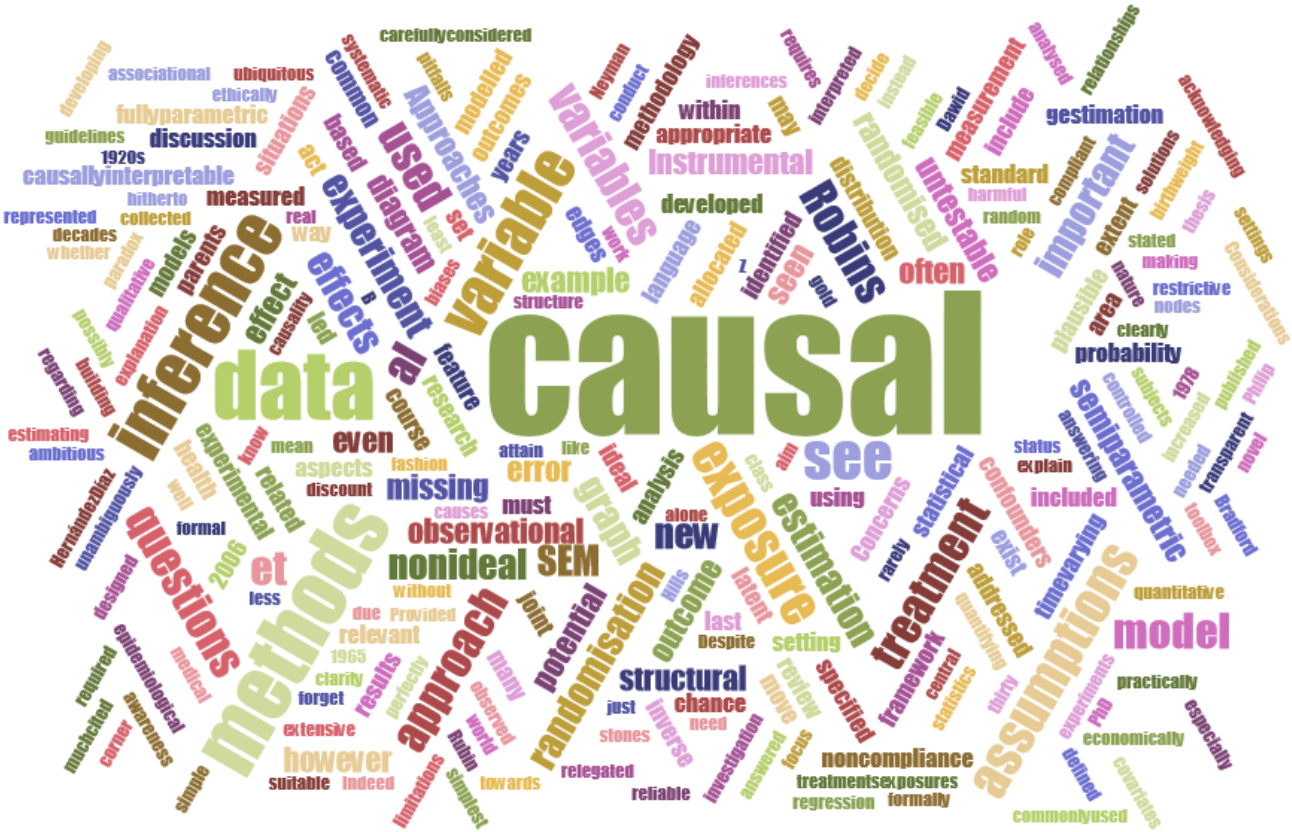
Potential Outcomes

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

Causal Effects

- Average Treatment Effect: **E[Y(1) - Y(0)]**.

COUNTERFACTUAL FRAMEWORK



Counterfactual framework: Causal Inference

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.

Counterfactual framework: Causal Inference

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**.

Counterfactual framework: Causal Inference

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**.

Counterfactual framework: Causal Inference

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**.

Justification

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.

Justification

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)

Justification

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.)

The challenge of IE

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 effect

Causal effects in the real world

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

Causal effect

Causal effects in the real world

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

The Outcomes for the treated and control individuals are:

Causal effect

Causal effects in the real world

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

The Outcomes for the treated and control individuals are:
$Y_i(1) = Y_i(A = 1)$ for some treatment variable A (Treated).

Causal effect

Causal effects in the real world

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

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 effect

Causal effects in the real world

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

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 effect

Causal effects in an ideal world

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

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 $Y_i(1) = Y_i(A = 1)$ (Treated)

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 $Y_i(1) = Y_i(A = 1)$ (Treated)

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

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 $Y_i(1) = Y_i(A = 1)$ (Treated)

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

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

However, we only observe:

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

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**.

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

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

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

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

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)

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

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)

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

Causal effect in OBSERVATIONAL study

ASSUMPTIONS to consider a CAUSAL EFFECT

- Rosebaum & Rubin, 1983: **The Ignorable Treatment Assignment Assumption** (Unconfoundeness, 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

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

Causal effect in an EXPERIMENTAL study

The solution to the fundamental problem of Causal inference

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

Causal effect in an EXPERIMENTAL study

The solution to the fundamental problem of Causal inference

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

RANDOMIZATION makes the **ATE** UNBIASED:

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

Causal effect in an EXPERIMENTAL study

The solution to the fundamental problem of Causal inference

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

RANDOMIZATION makes the **ATE** UNBIASED:

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

Causal effect in an EXPERIMENTAL study

The solution to the fundamental problem of Causal inference

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

RANDOMIZATION makes the **ATE** UNBIASED:

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

The solution to the fundamental problem of Causal inference

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

Causal effects in OBSERVATIONAL studies

When randomization is unethical or infeasible

Causal effect is biased (B):

$$\textbf{ATE} + \textbf{B}$$

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):

$$\textbf{ATE} + \textbf{B}$$

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):

$$\textbf{ATE} + \textbf{B}$$

Type of bias

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

Causal effect in OBSERVATIONAL studies

ASSUMPTIONS to consider a CAUSAL EFFECT

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

Causal effect in OBSERVATIONAL studies

ASSUMPTIONS to consider a CAUSAL EFFECT

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

Causal effect in OBSERVATIONAL studies

ASSUMPTIONS to consider a CAUSAL EFFECT

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

Causal effect in OBSERVATIONAL studies

ASSUMPTIONS to consider a CAUSAL EFFECT

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

Causal effect in OBSERVATIONAL studies

ASSUMPTIONS to consider a CAUSAL EFFECT

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

Causal effect in OBSERVATIONAL studies

ASSUMPTIONS to consider a CAUSAL EFFECT

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

Causal effects in OBSERVATIONAL studies

How to use the Neyman-Rubin Counterfactual framework?

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

Causal effects in OBSERVATIONAL studies

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.

Causal effects in OBSERVATIONAL studies

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.

Causal effects in OBSERVATIONAL studies

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

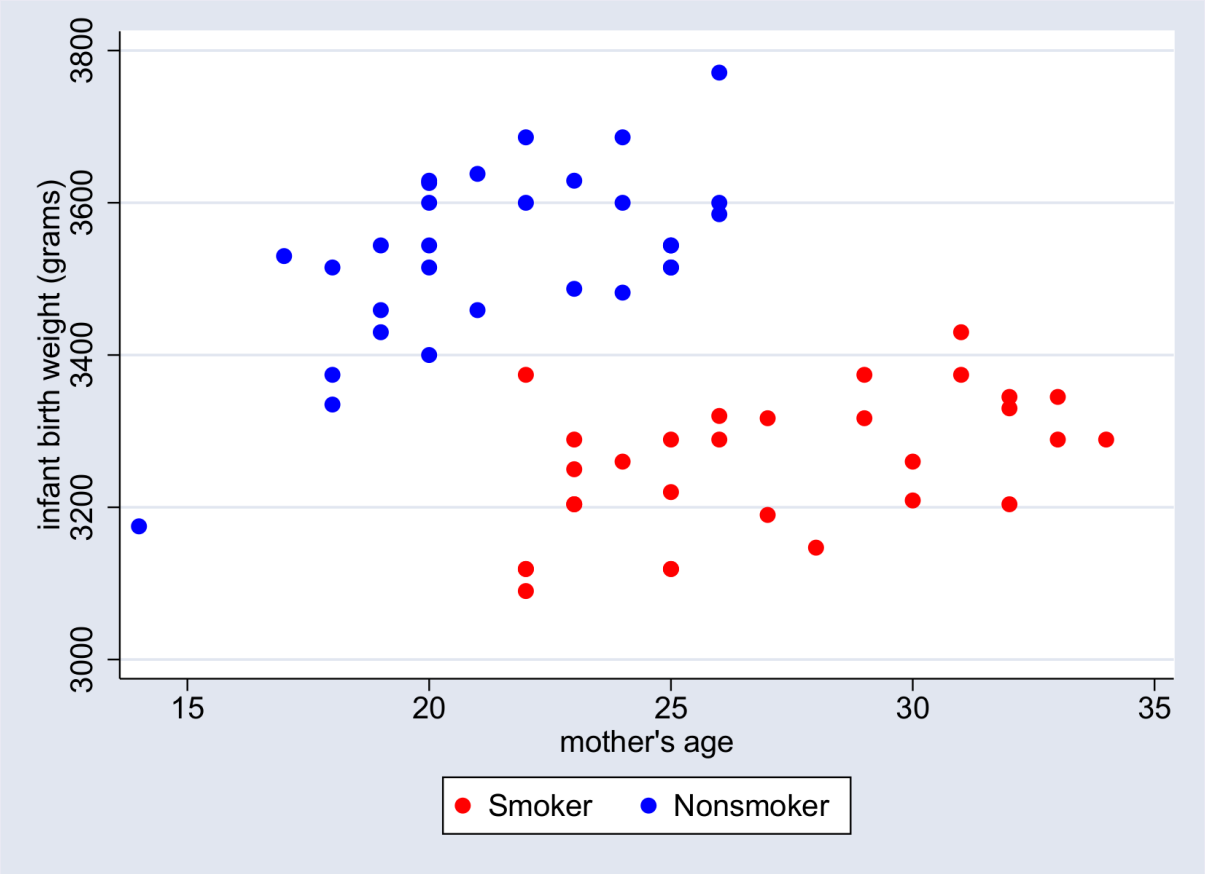
In an ideal world, we would see this:

$Unit_i$	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

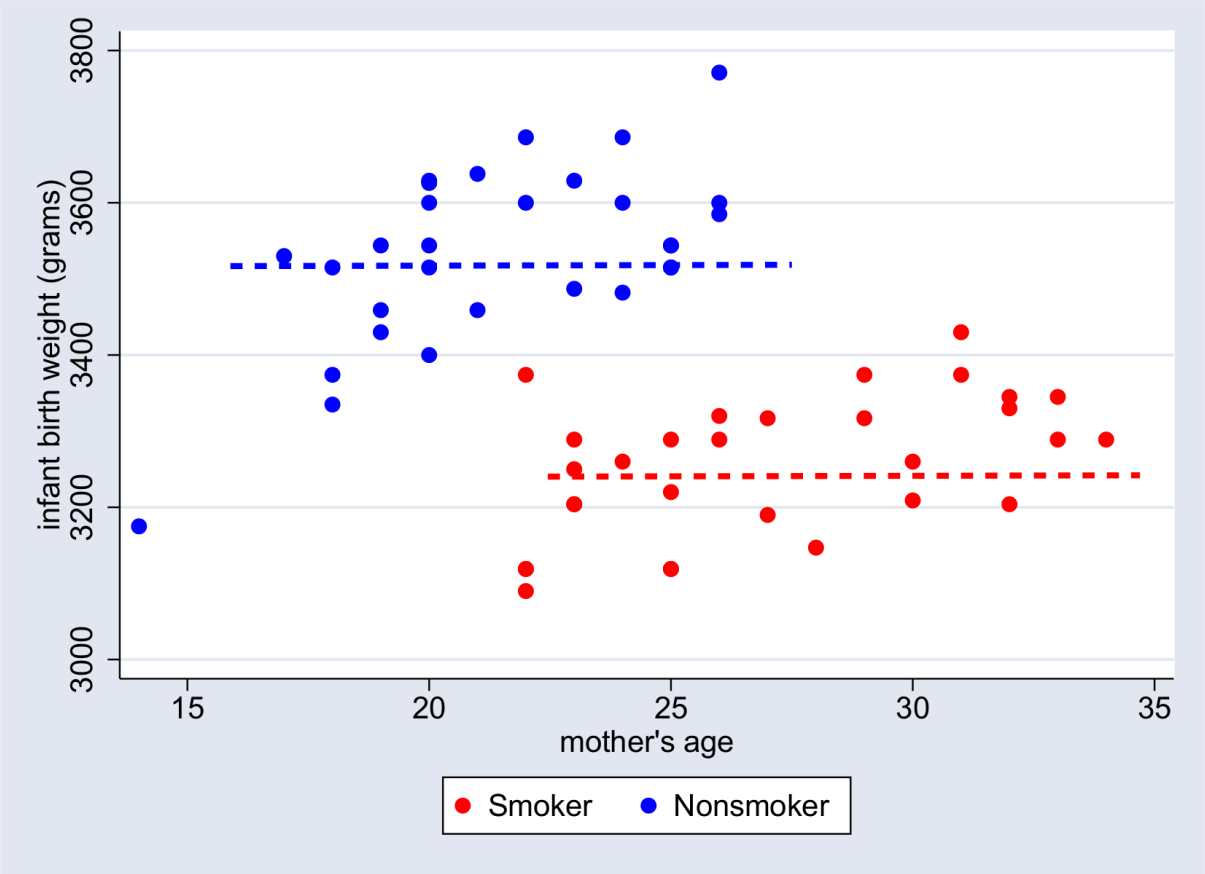
In the real world, this is what we see:



Causal effect

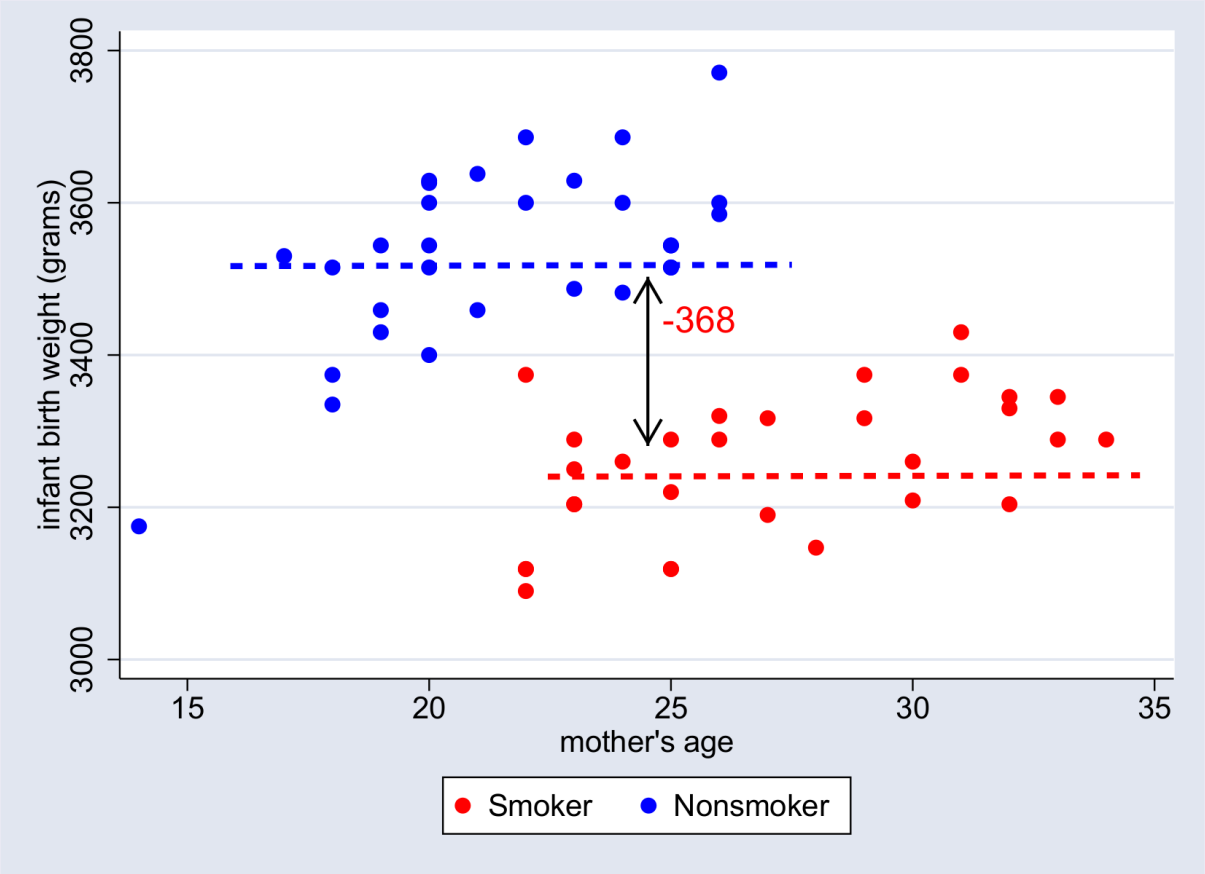
The fundamental problem of Causal inference

In the real world, this is what we see:



The fundamental problem of Causal inference

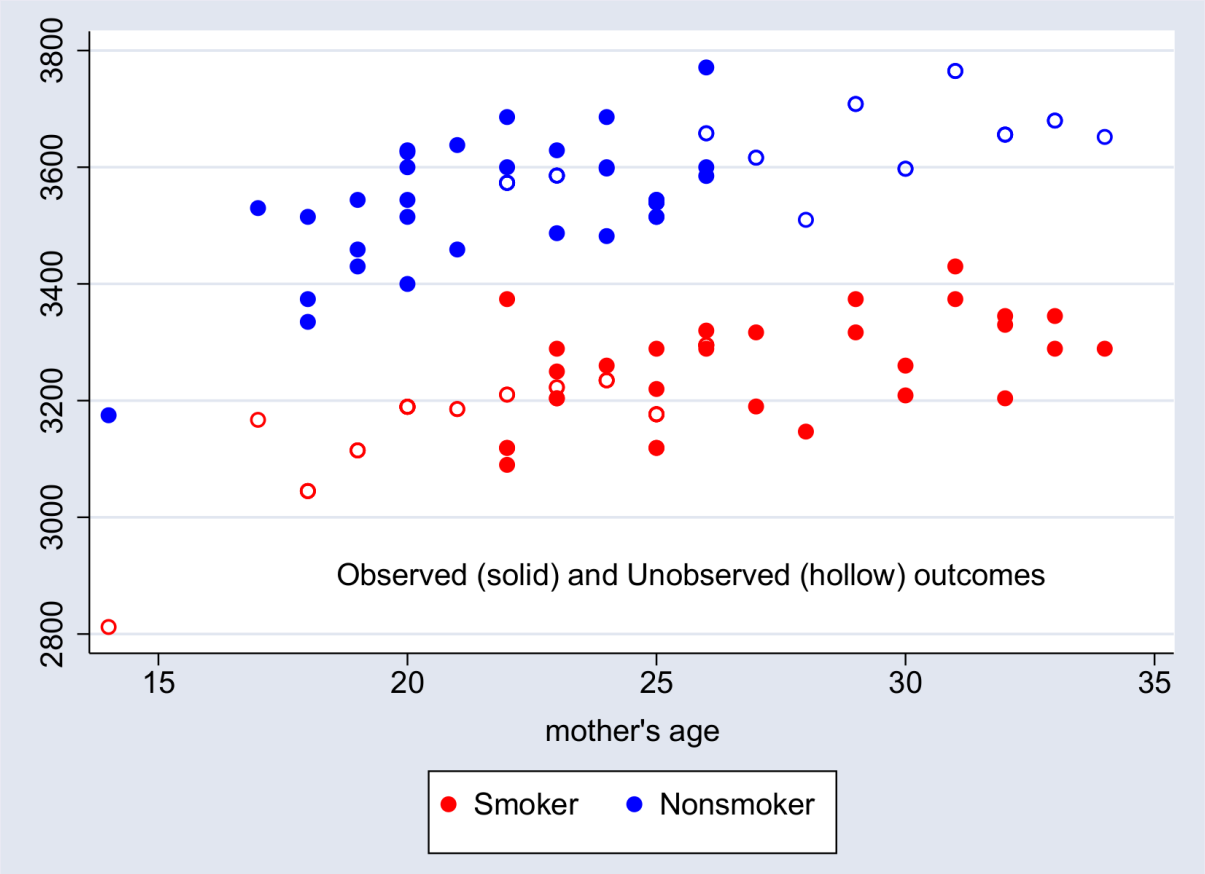
In the real world, this is what we see:



Causal effect

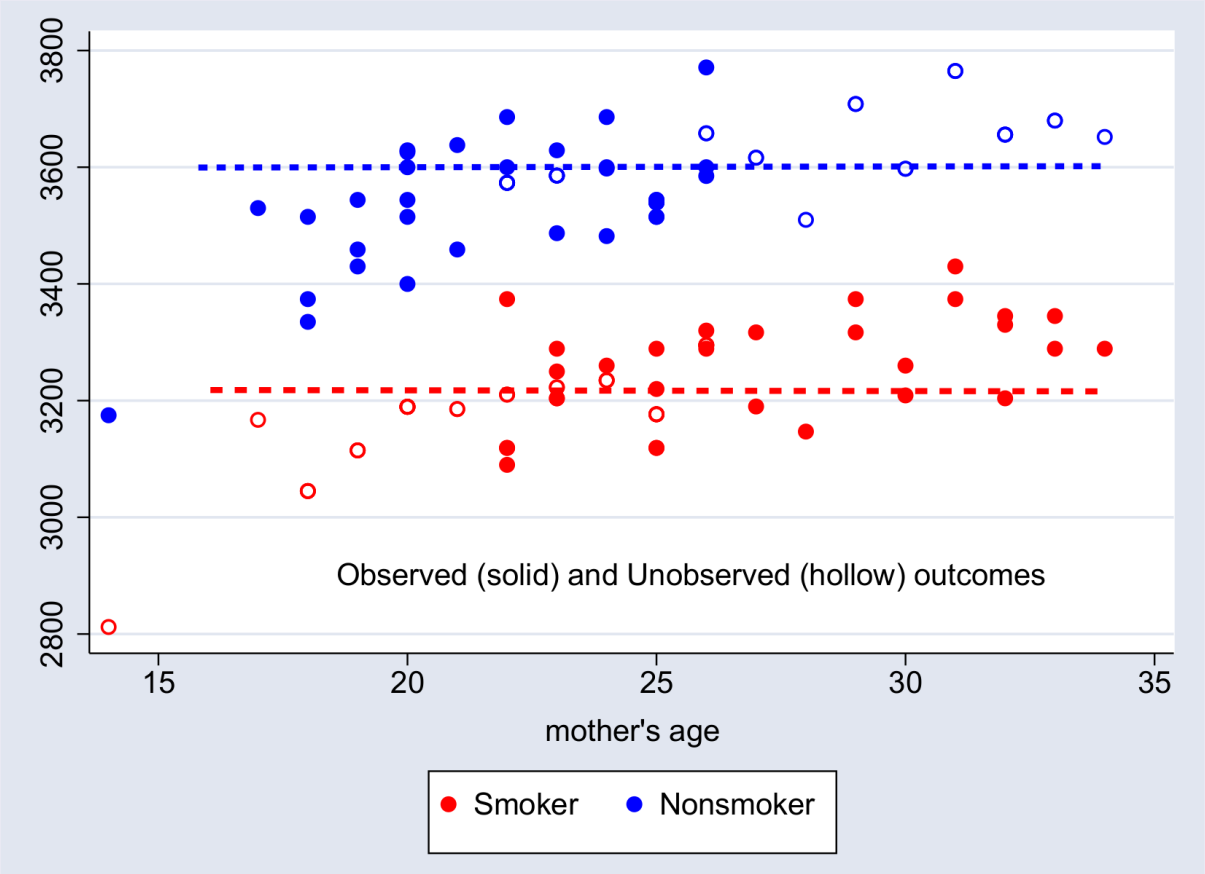
The fundamental problem of Causal inference

In an ideal world, this is what we see:



The fundamental problem of Causal inference

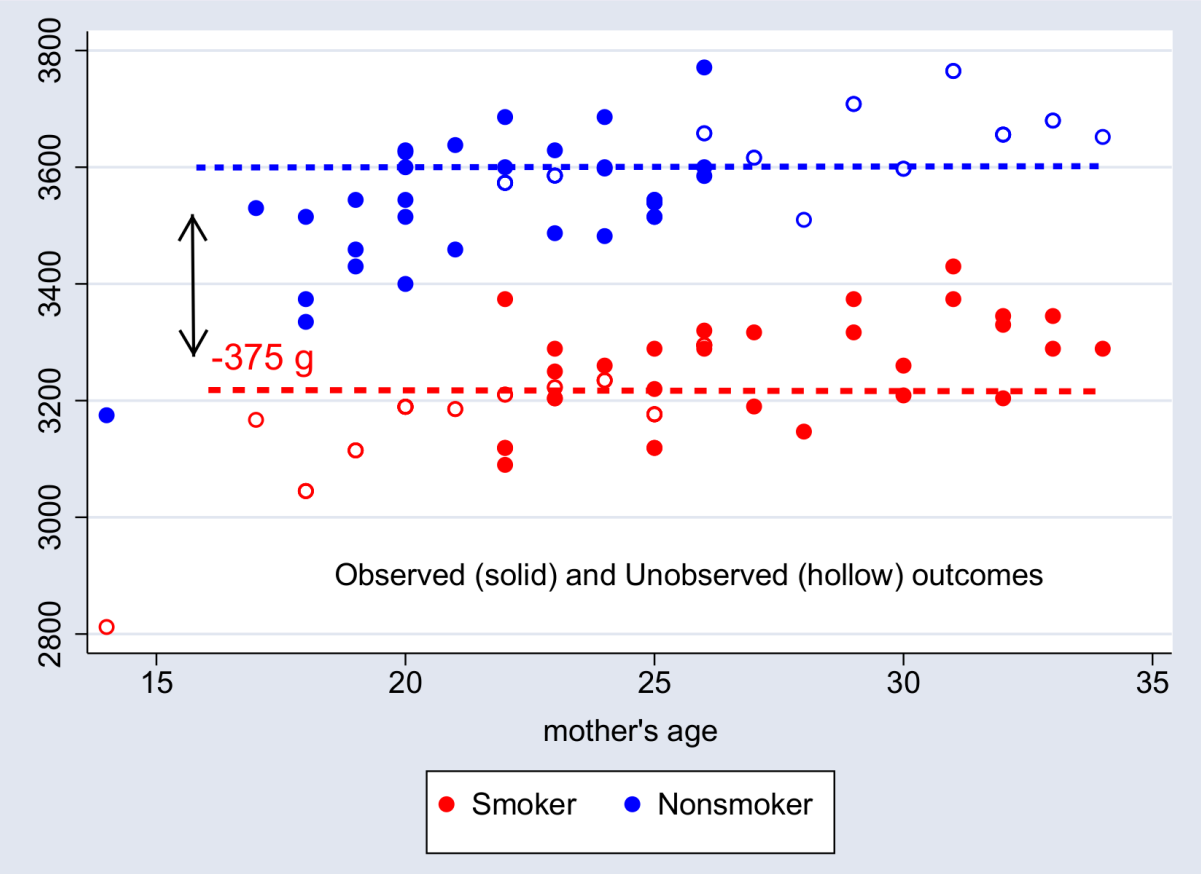
In an ideal world, this is what we see:



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 $\mathbf{Y_i(A = 0) = Y_i(0)}$.

Controls $Y_i(0)$ would have been treated $\mathbf{Y_i(A = 1) = Y_i(1)}$.

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

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

POSITIVITY

POSITIVITY: $\mathbf{P(A = a \mid W) > 0}$ for all a, W

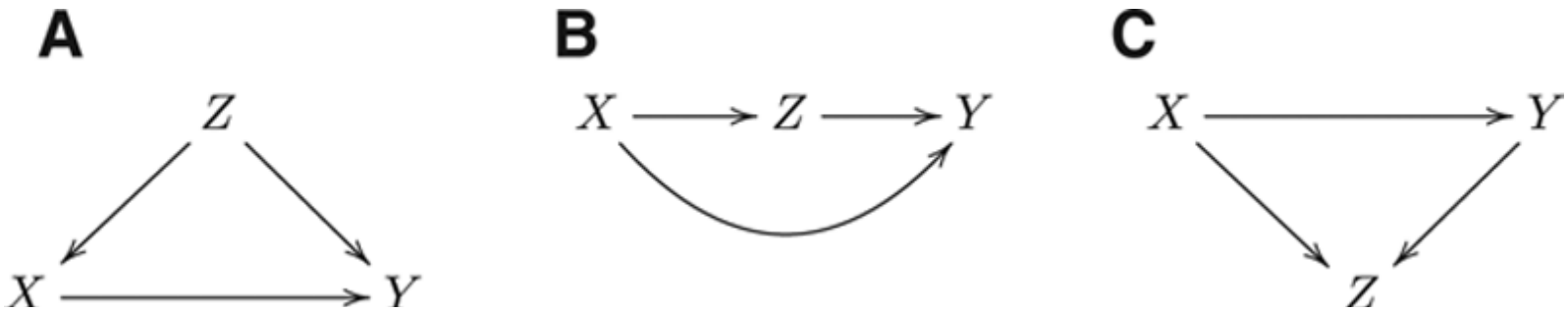
SUTVA

- We have assumed that there is only on 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



Ignorability: DAGs

Better control of confounding: Ignorability

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

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

Ignorability: DAGs

Better control of confounding: Ignorability

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

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

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

Ignorability: DAGs

Better control of confounding: Ignorability

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

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

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

Ignorability: DAGs

Better control of confounding: Ignorability

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

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

- 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.

Ignorability: DAGs

Better control of confounding: Ignorability

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

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

- 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).

Ignorability: DAGs

Better control of confounding: Ignorability

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

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

- 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).

DAGs

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.

DAGs

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 ($Z <- W -> Y$) .

DAGs

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 ($Z <- W -> Y$) .
- 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.

DAGs

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 ($Z <- W -> Y$) .
- 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.

Confounding

Properties of a confounder

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

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).

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.

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.

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.
- Must be not affected by (common cause of) the exposure or (common cause of disease): collider.

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.
- Must be not affected by (common cause of) the exposure or (common cause of disease): collider.
- Cannot be an intermediary cause: mediator

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.
- Must be not affected by (common cause of) the exposure or (common cause of disease): collider.
- Cannot be an intermediary cause: mediator

Colliders

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

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

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.

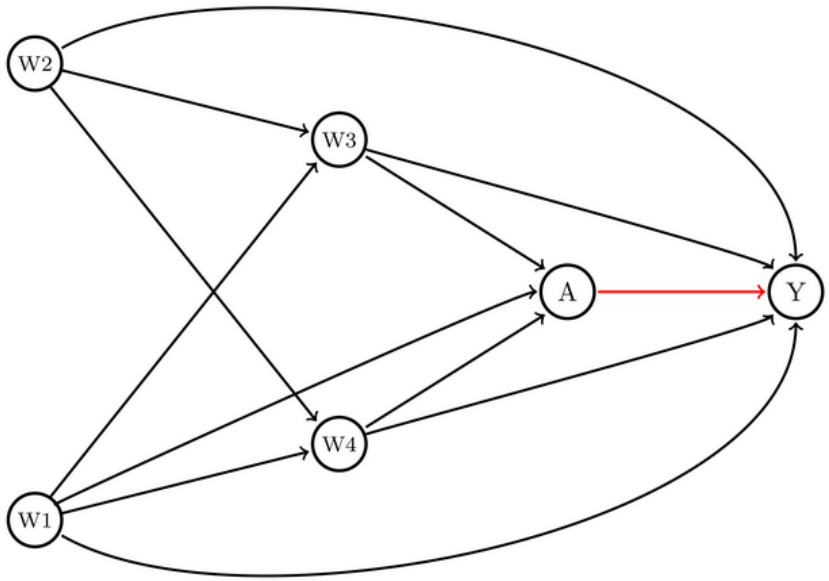
- Deconstructing paradoxical effects in medical literature: Luque-Fernandez MA et al. Deconstructing the **smoking-preeclampsia paradox** through a counterfactual framework. Eur J Epidemiol. 2016;31:613-623 (https://www.ncbi.nlm.nih.gov/pubmed/26975379).

DAG example: back door paths

Direct Acyclic Graph (DAG)

Under conditional exchangeability: $Y(0), Y(1) \perp A|W$

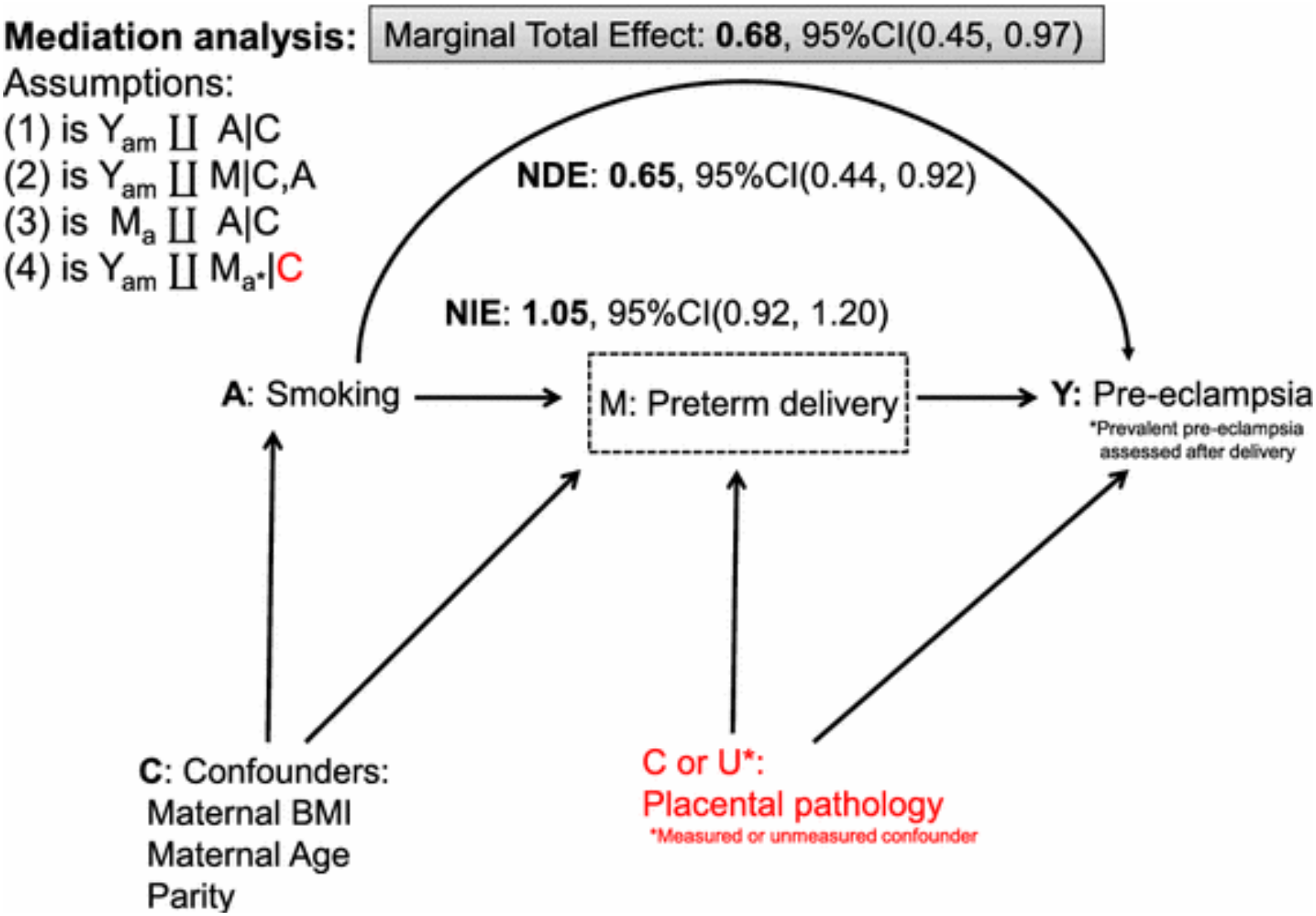
$$\text{ATE} = E[E(Y|A=1;W) - E(Y|A=0;W)]$$



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

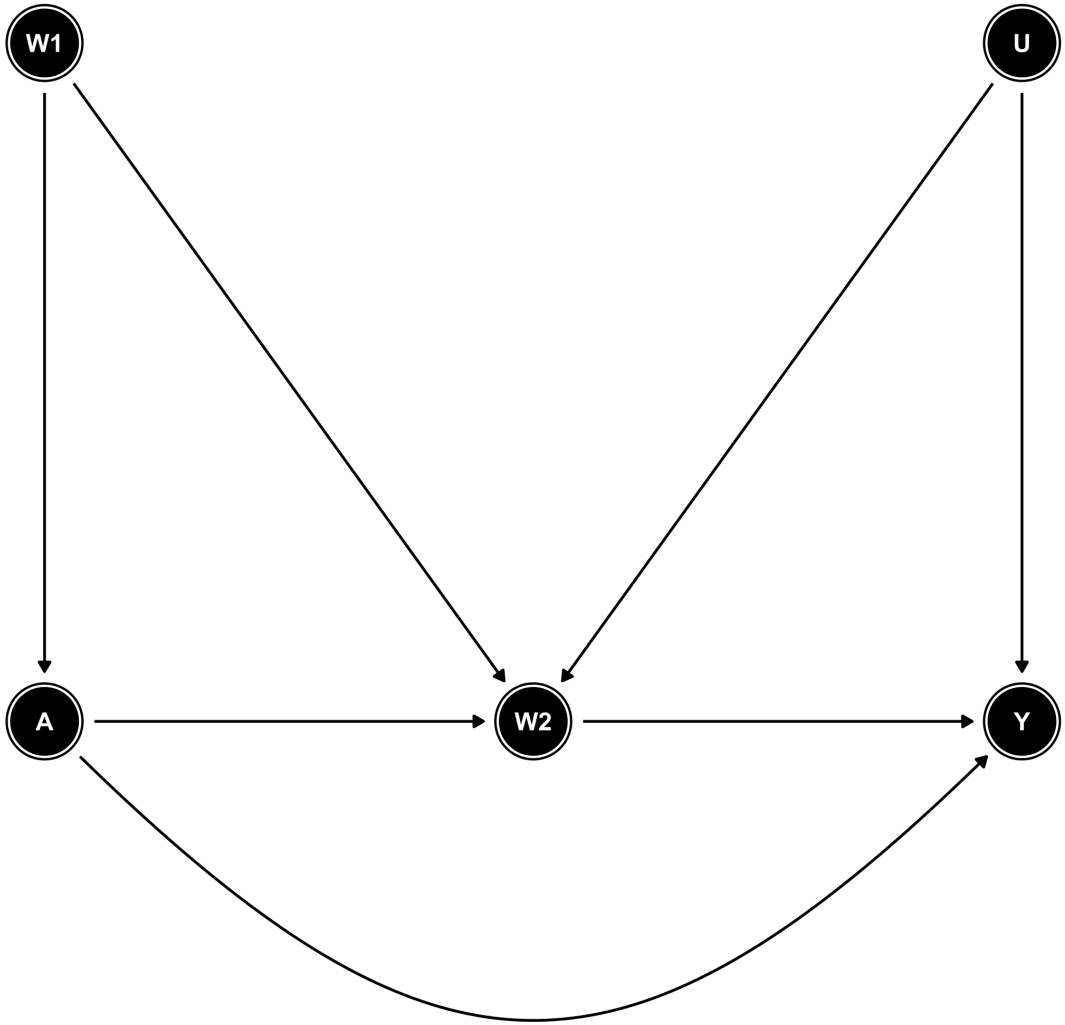
[5] Luque-Fernandez MA et al. Data-Adaptive Estimation for Double-Robust Methods in Population-Based Cancer Epidemiology: Risk differences for lung cancer mortality by emergency presentation (2017). AJE. https://academic.oup.com/aje/article/doi/10.1093/aje/kwx317/4110407

DAG example: mediation



[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

SWIG

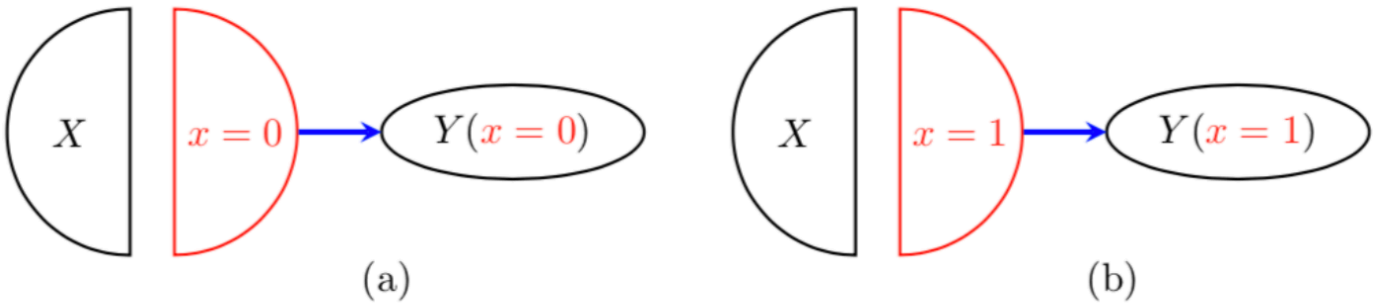


Figure 2: The single world intervention graphs (SWIGs) resulting from splitting node X in the graph in Figure 1(a), and intervening to set a particular value. (a) the SWIG $\mathcal{G}(x=0)$ corresponding to setting X to 0; (b) $\mathcal{G}(x=1)$ given by setting X to 1.

[8] Robins J. (2013) Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality

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
```

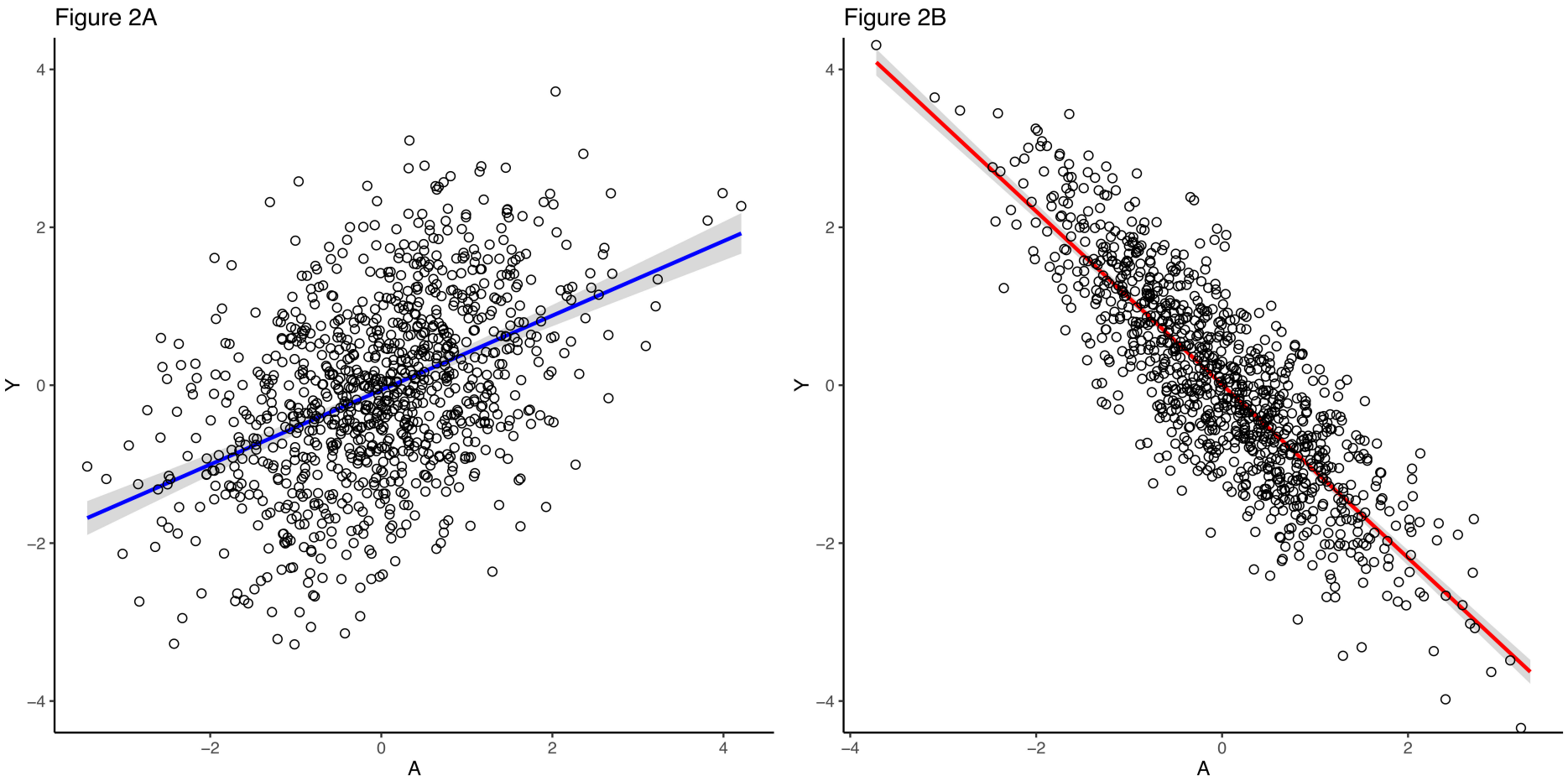
Collider and confounding effects

	Dependent variable (Y)			
	W (confounder)		C (collider)	
	Unadjusted β (SE)	Adjusted β (SE)	Unadjusted β (SE)	Adjusted β (SE)
	(Fit 1)	(Fit 2)	(Fit 3)	(Fit 4)
A	**0.471**	**0.289**	**0.326**	**-0.416**
	(-0.030)	(-0.032)	(-0.031)	(-0.035)
W		0.425		0.491
		(-0.035)		(-0.018)
Intercept	-0.061	-0.06	0.01	0.035
	(-0.033)	(-0.031)	(-0.031)	(-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

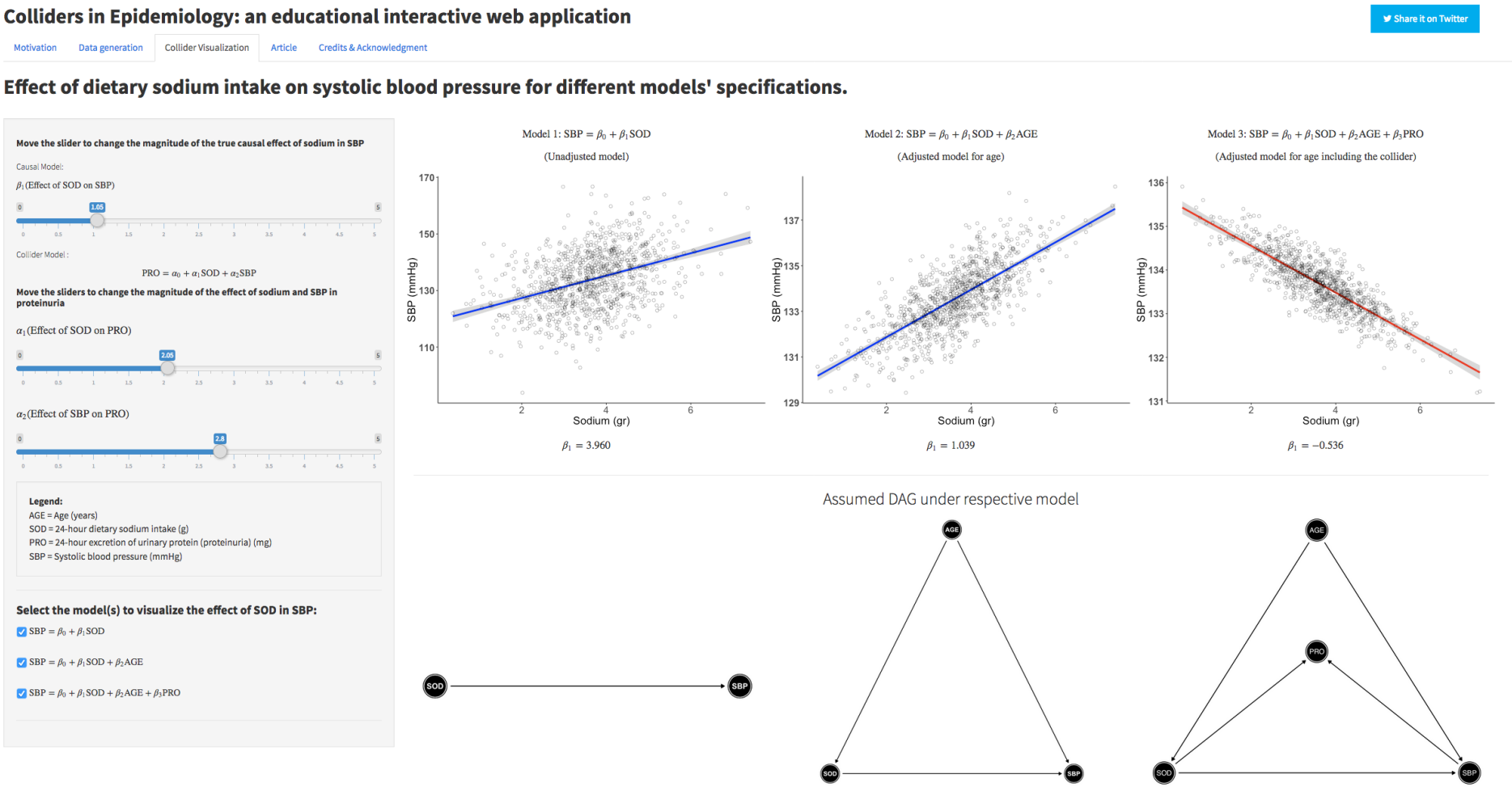
Display Linear Fit: models (fit2) and (fit4)



Collider Effect

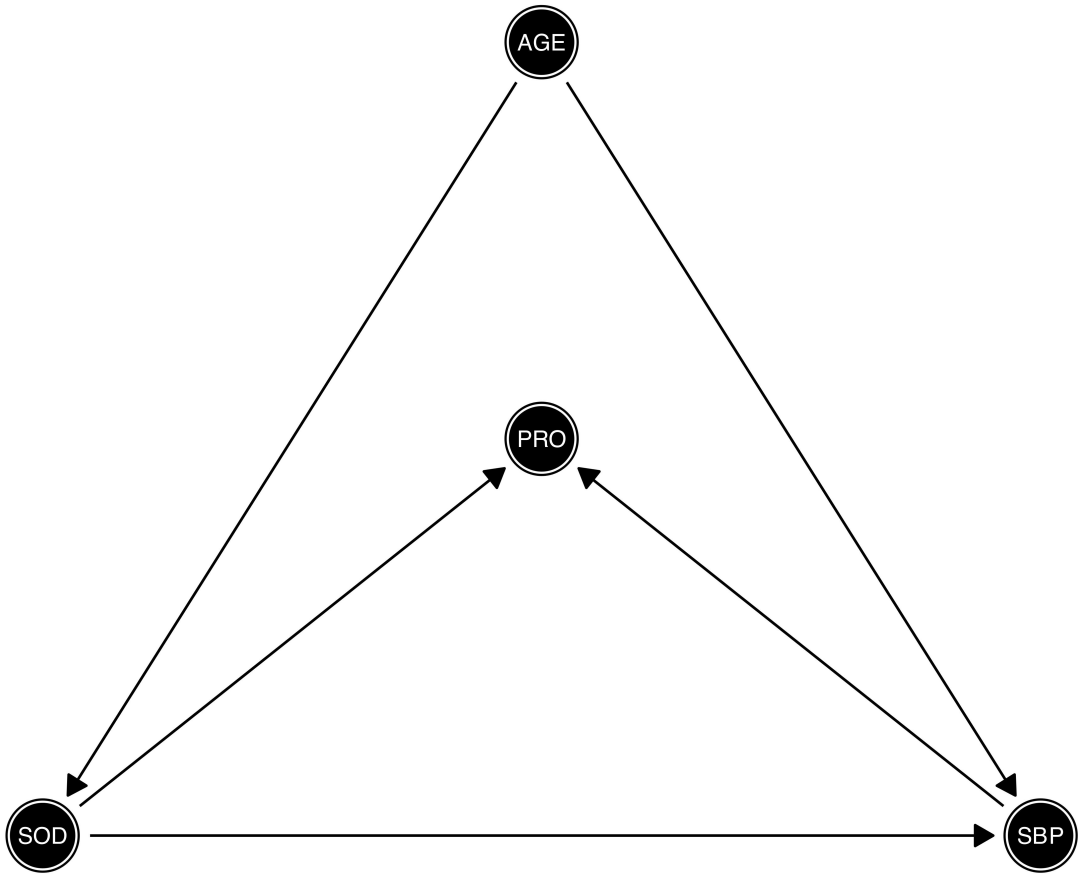
Luque-Fernandez et al.(2018) Educational Note: Paradoxical Collider Effect in the Analysis of Non-Communicable Disease Epidemiological Data: a reproducible illustration and web application https://arxiv.org/abs/1809.07111

Shiny web application



Colliders in Epidemiology: an educational interactive 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

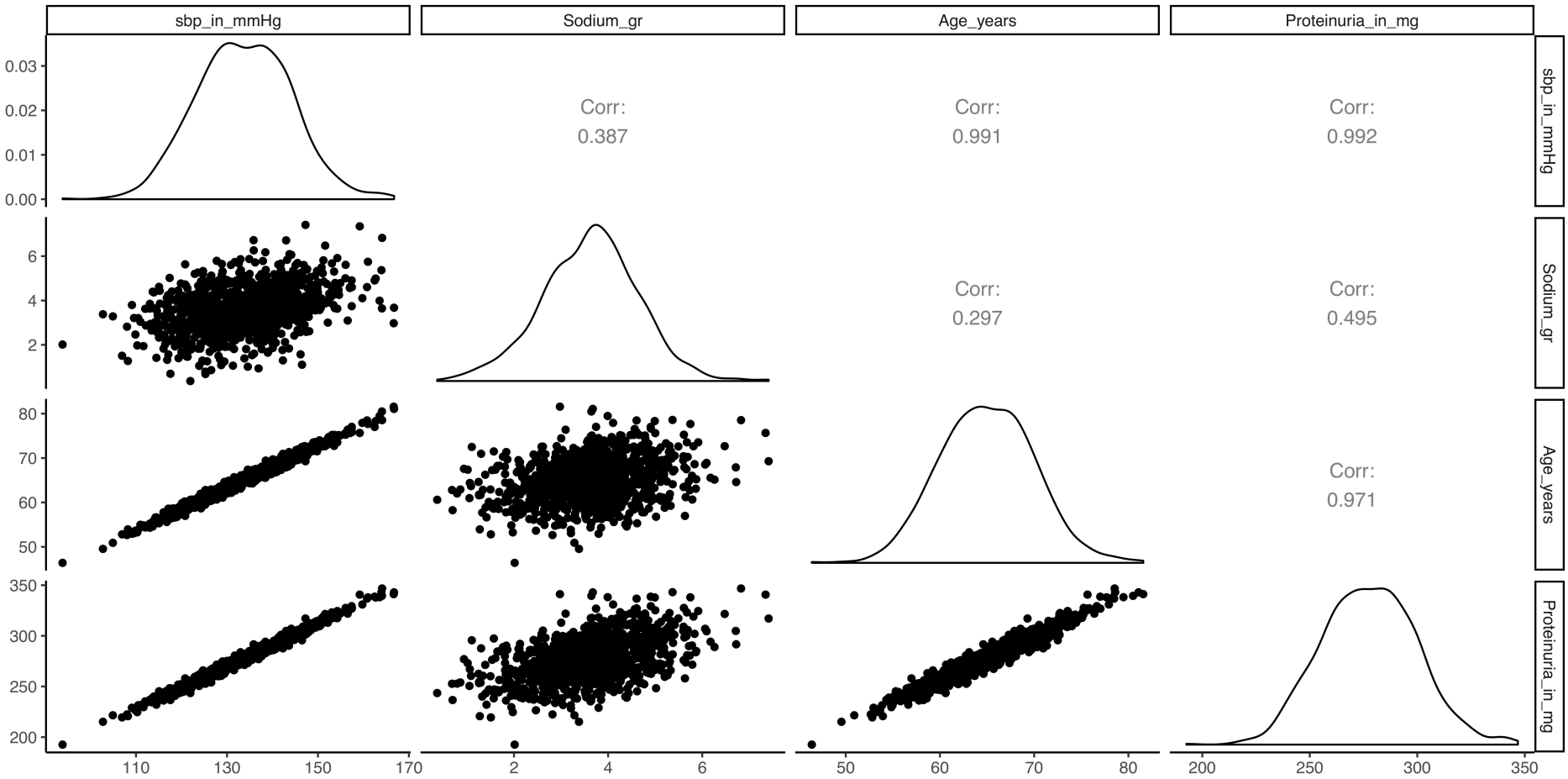
```
generateData <- function(n, seed){
set.seed(seed)
Age_years <- rnorm(n, 65, 5)
Sodium_gr <- Age_years / 18 + rnorm(n)
sbp_in_mmHg <- 1.05 * Sodium_gr + 2.00 * Age_years + rnorm(
hypertension <- ifelse(sbp_in_mmHg>140,1,0)
Proteinuria_in_mg <- 2.00*sbp_in_mmHg + 2.80*Sodium_gr + rn
data.frame(sbp_in_mmHg,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

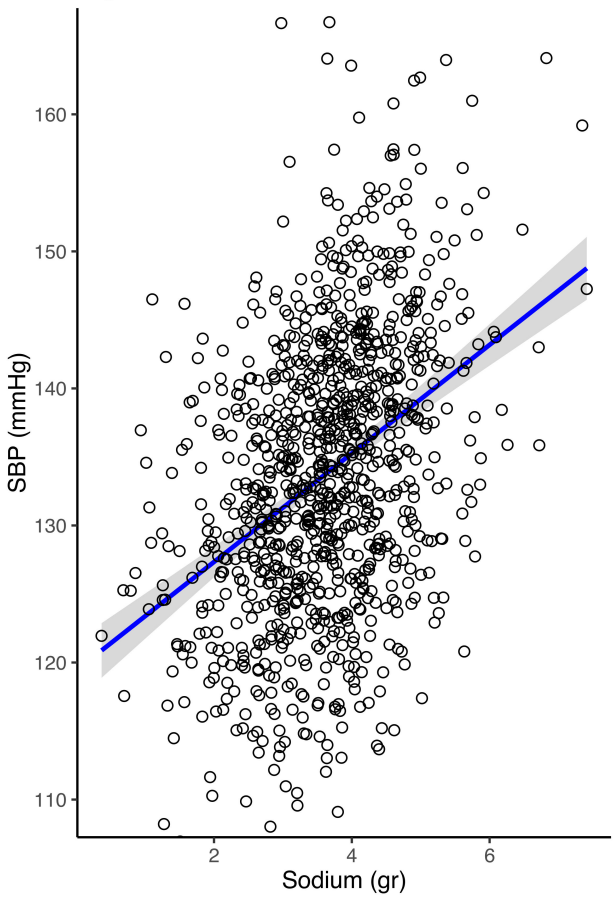


Figure 4B

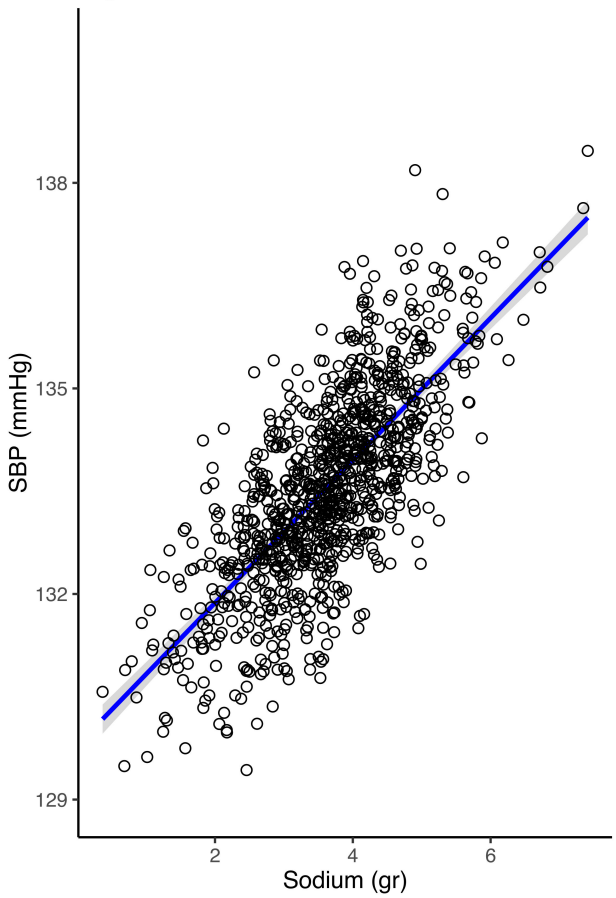
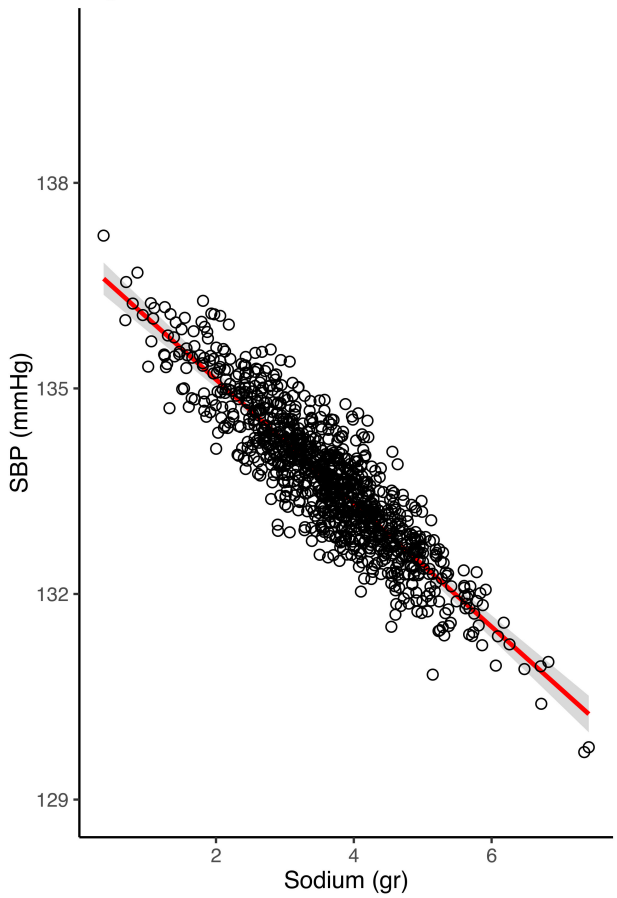


Figure 4C



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 (SE)	Bivariate (SE)	Multivariate (SE)
True effect of Sodium in gr: 1.05			
Sodium in gr	**3.960**	**1.039**	**-0.902**
	(0.298)	(0.032)	(0.036)
Age in years		2.004	0.416
		(0.007)	(0.027)
Proteinuria in mg			0.396
			(0.007)
Intercept	119.420	-0.311	-0.091
	(1.122)	(0.407)	(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!

Miguel Angel Luque-Fernandez

miguel.luque.easp@juntadeandalucia.es @watzilei

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