

Clinical Epidemiology in the Era of Big Data and Data Science: New Opportunities

Miguel Angel Luque-Fernandez, PhD

Assistant Professor of Epidemiology
Faculty of Epidemiology and Population Health
Department of Non-communicable Disease Epidemiology
Cancer Survival Group

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

November 2, 2017



Public Health as Scientific Discipline: subdisciplines

The screenshot shows the official website of the Harvard T.H. Chan School of Public Health. The header includes the school's logo, a search bar, and navigation links for 'Most Visited' (38th Annual Conference), 'Home - Research Part...', 'Surveillance Research ...', 'Tutoriales - Big Data y...', and 'A Primer in Econ...'. Below the header is a red banner with links for 'INFORMATION FOR: Prospective Students', 'Current Students', 'Alumni', 'Faculty & Staff', and 'Friends & Supporters'. The main navigation menu includes 'ABOUT', 'FACULTY & RESEARCH', 'ADMISSIONS & AID', 'ACADEMICS', 'EXECUTIVE/CONTINUING ED', and 'NEWS'. The current page is 'Departments', indicated by the bold text in the menu and the breadcrumb trail 'Home > Departments'. On the left, a sidebar lists various academic units and programs. The main content area features a heading 'Academic Departments, Divisions and Centers' and a section titled 'Academic Departments' with a list of ten sub-disciplines, each preceded by a downward arrow icon.

Departments | Harvard T.H. Chan School of Public Health

Most Visited 38th Annual Conference Home - Research Part... Surveillance Research ... Tutoriales - Big Data y... A Primer in Econ...

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

Home > Departments

DEPARTMENTS

Search this section

Home

Academic Departments

Divisions

Research Centers

Flagship Initiatives

International Research

Research Administration and Support

Degree Programs

Fellowships and Residencies

Summer Programs

Continuing Professional Education

Interdisciplinary Concentrations

Academic Calendar

Harvard Chan Viewbook

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 subdiscipline: 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
- Pharmacopidemiology
- 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 causes 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 improve health and reduce health inequities around the world.

Innovative Educational Programs: Onsite and Online

 MEDICINE

Data Science Initiative

co-directors of newly launch x +

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

Visited 38th Annual Conferenc... Home - Research Part... Surveillance Research ... Tutoriales - Big Data y... A Primer in Economet... GESTIÓN DE PROYECTO... DeepL Translator

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

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

Data Science Programmes

A screenshot of a web browser window. The address bar shows 'data.harvard.edu/education'. The page content is about data science programs for graduate students, featuring sections on Master's Degree Programs in Data Science, PhD Programs in Data Science, and Doctoral Degree Secondary Field Opportunities. The navigation menu includes links for HOME, ABOUT, NEWS, COMMUNITY, EDUCATION, EVENTS, and APPLY. The top right corner has icons for search, star, folder, download, and globe.

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

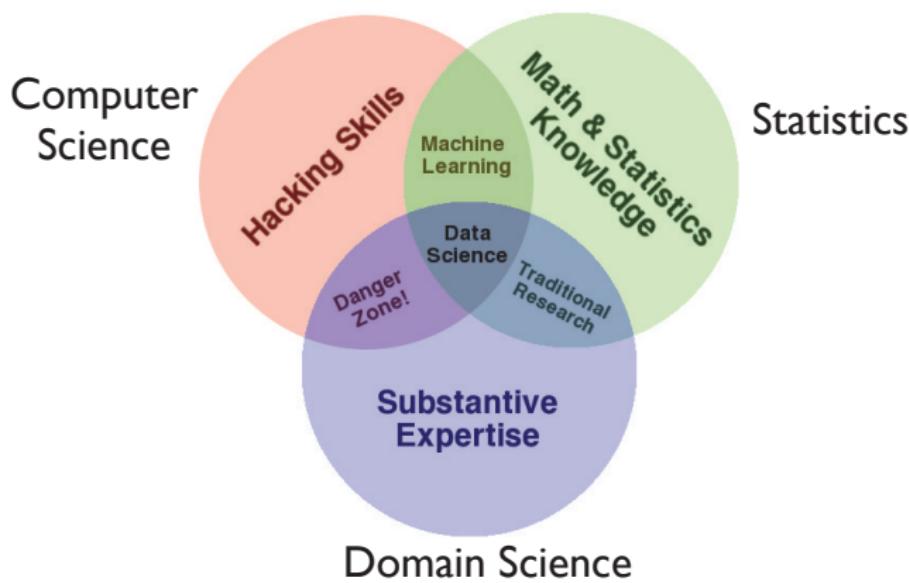
[PhD in Bioinformatics and Integrative Genomics \(BIG\)](#)

Doctoral Degree Secondary Field Opportunities

NDON
OOL OF
IENE
OPICAL
DICINE



Data Science

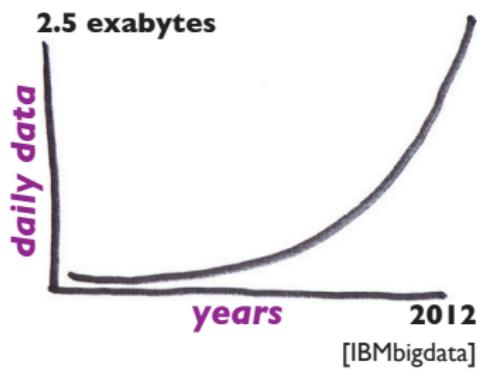


Drew Conway

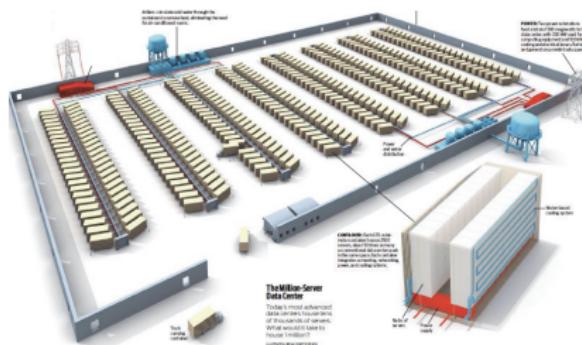


Data Science and Big Data: Volume

Big Data



Commodity Computing



Michael Franklin, UC Berkeley



Smarter Devices

Smart devices



GarageBand '08

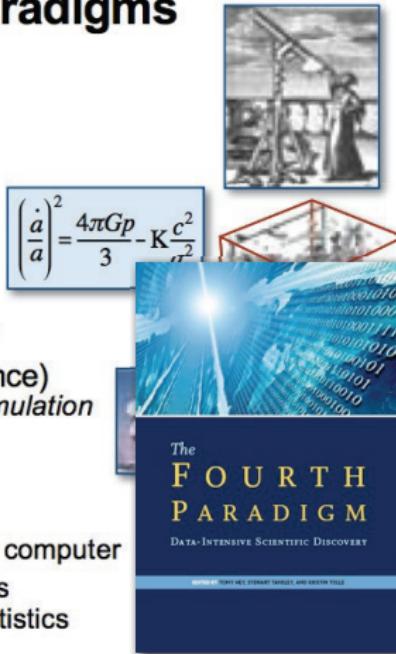


Michael Franklin, UC Berkeley



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”

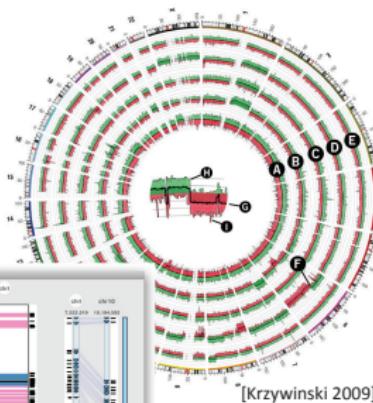
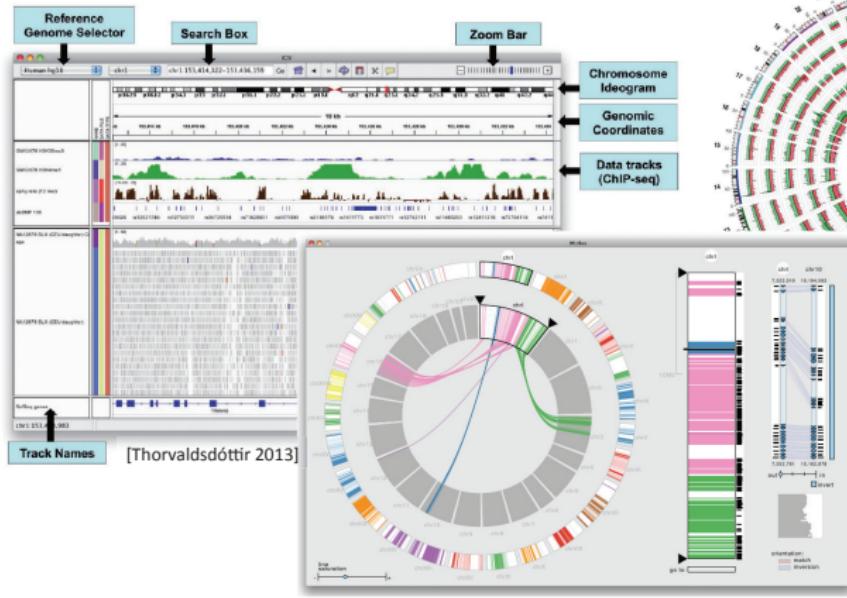
McKinsey Global Institute

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

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



Genome Visualization



[Meyer 2009]



So, how about Epidemiology?

electronic charting

health record

digital

patient

nursing

doctor

con



So, how about Epidemiology?



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 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

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

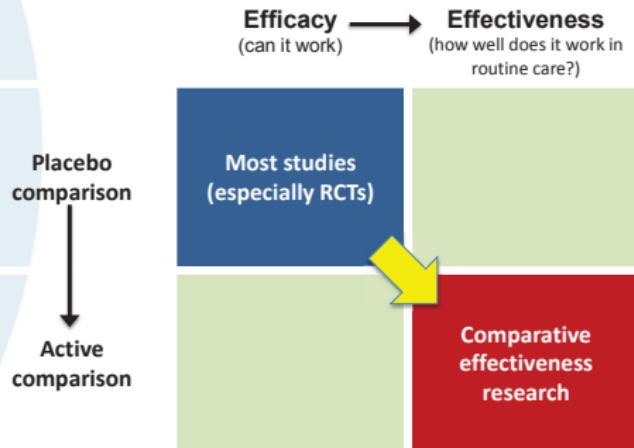
...is the generation and synthesis of evidence that
**compares the benefits and harms of alternative
methods to prevent, diagnose, treat, and
monitor a clinical condition or to improve the
delivery of care.**

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



What CER seeks to do

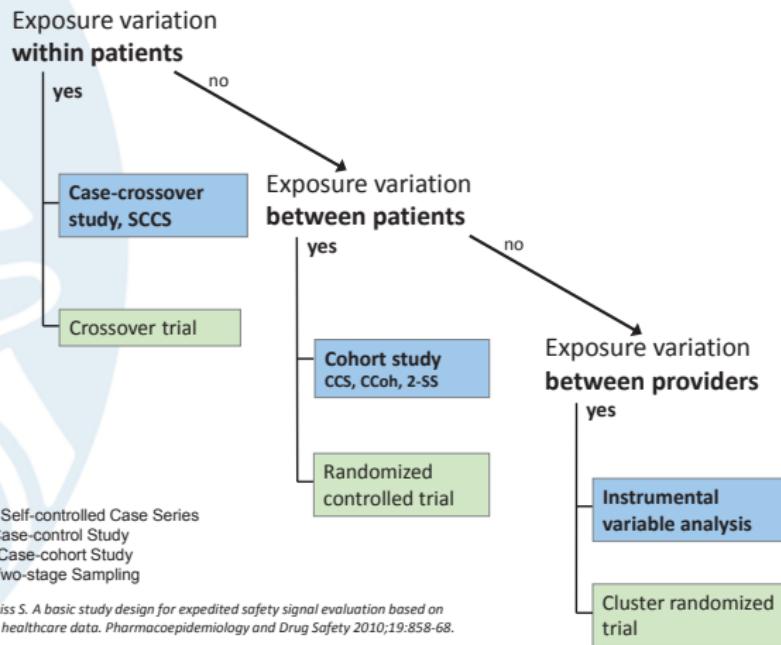
	TYPICAL RCTs	NEEDS OF DECISION MAKERS
Comparator	Placebo or usual care	Active
Patient population	Highly selected	Representative of typical practice
Outcome measures	Surrogate	Patient centered
Follow-up time	Short	Long
Cost	High	Moderate
Speed	Slow	Faster

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



CER is about New Epidemiological Methods

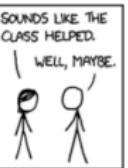
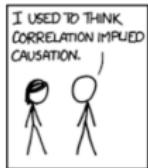
Design choice: Source of exposure variation



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



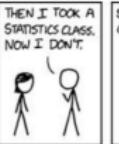
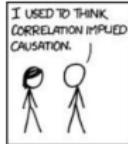
CAUSAL INFERENCE IN STATISTICS

A Primer

Judea Pearl
Madelyn Glymour
Nicholas P. Jewell

WILEY

Please go to www.wiley.com/go/causal-inference



A Research Agenda on Causal Inference

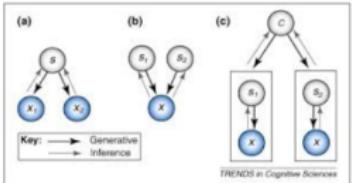
- Many problems in causal inference are not prediction and causal explanation.
- Existing M-estimators for causal selection are not able to directly apply to the estimation of causal parameters.
- Inference may need new methods.

Promises:
• Formally model the distinction between causal and non-causal parts of the model and test them against observational data.
• Assess Amy, Inverse and Unmeasured Variables.

- Develop new estimation methods that combine the strengths of different causal approaches.
- Develop new approaches to assess validation of causal models without relying on validation experiments.
- Develop robustness measures for causal inference.
- Develop methods for causal discovery and causal analysis, drawing on CDS for causal inference.
- Large-scale causal databases with open variables.

Predictor

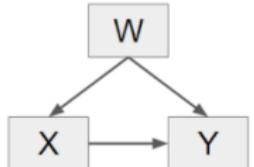
Child negative birth weight
weight at birth
days in hospital
first born
age
Mother
black
hispanic
white
smoking at birth
less than high school
high school graduate
some college
college graduate
smoked during pregnancy
not married
age at birth



TRENDS in Cognitive Sciences

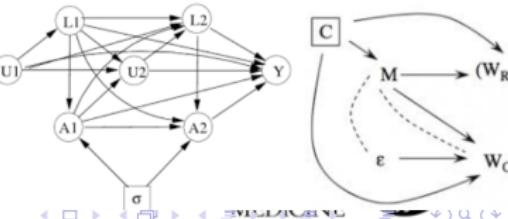
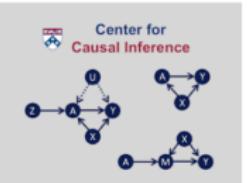
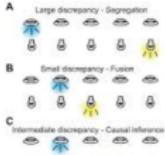
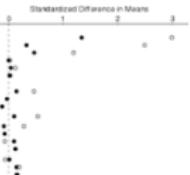
Causal Inference

- 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

causal inference - Google Search Deconstructing the smoking Deconstructing the smoking collapsibility odds ratio - Google 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 PROYECTO... 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

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.

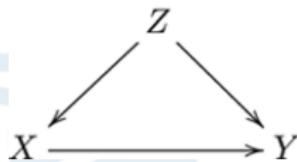
Causal Inference based on the Neyman-Rubin Potential Outcomes Framework (Rubin, 2011), first introduced in Social Science by Donal Rubin (Rubin, 1974) and later in Epidemiology and Biostatistics by James Robins (Greenland and Robins, 1986), has provided the theory and statistical methods needed to overcome 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.

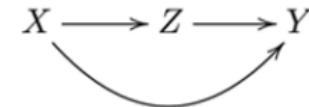
Causal effects are often formulated regarding comparisons of potential outcomes, as formalised by Rubin (Rubin, 2011). Let A denote a binary exposure, W a vector of potential confounders, and Y a binary outcome. Given A , each individual has a pair of potential outcomes: the outcome when exposed, denoted 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

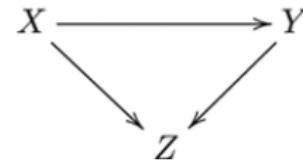
A



B



C

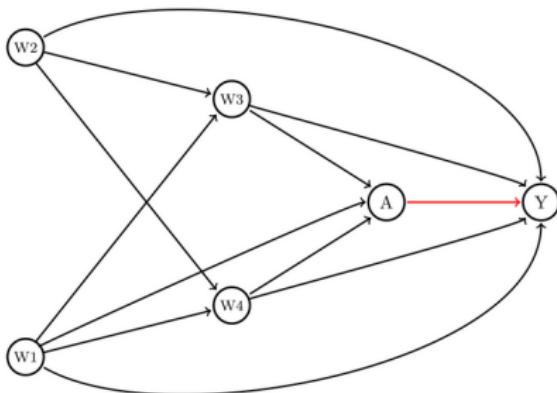


CER is about Causal Inference

Direct Acyclic Graph (DAG)

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

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



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

Source: 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>

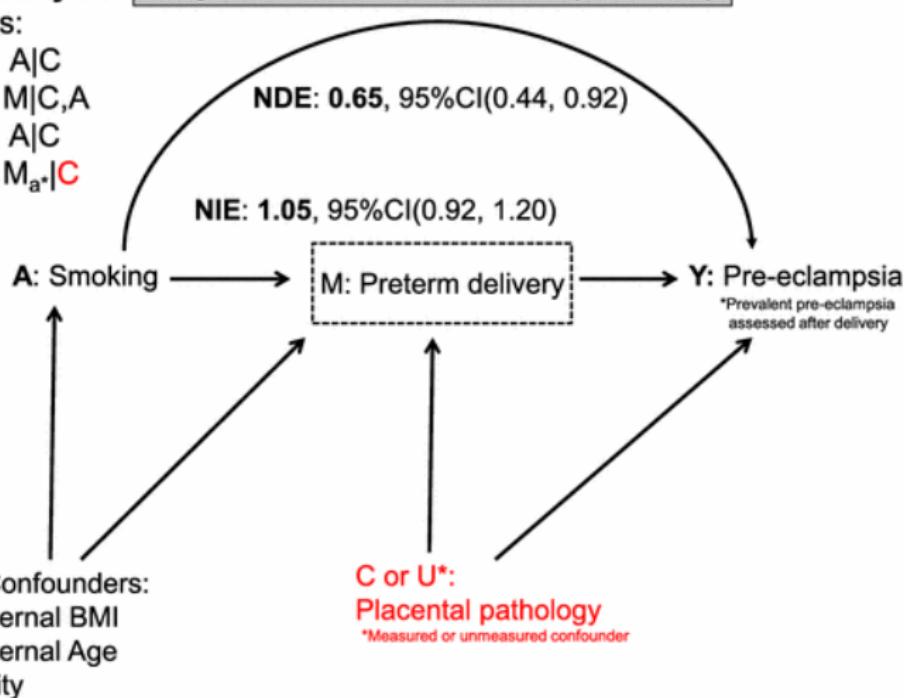


CER is about Causal Inference

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 $Y_{am} \perp\!\!\!\perp M_a|C$



Source: 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

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

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

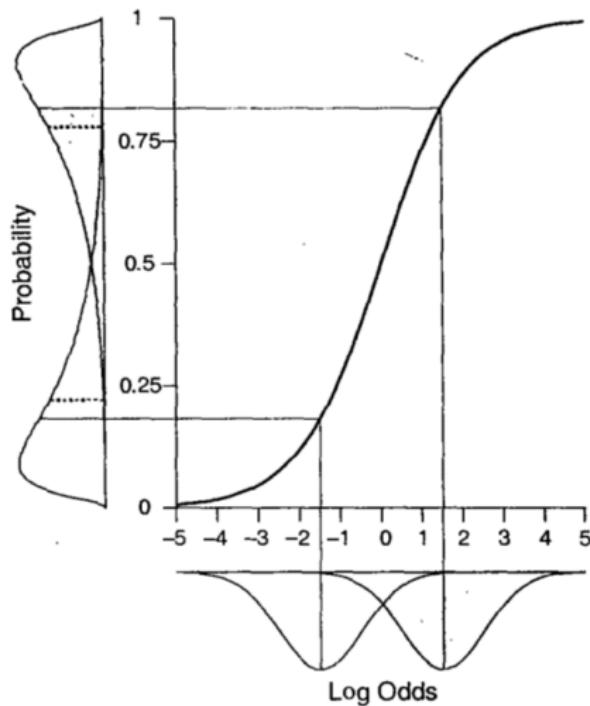
Most epidemiologists would not be surprised to find that the conditional odds ratio is different from the unadjusted marginal (over Z) odds ratio

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

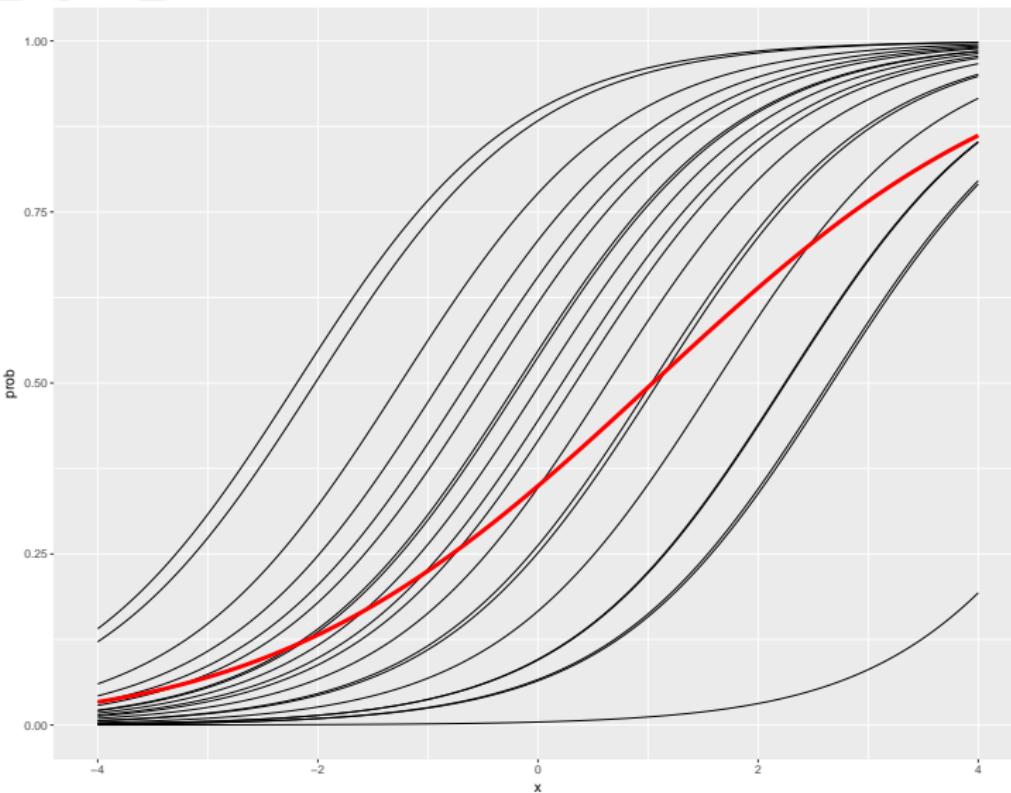
One explanation for a discrepancy between the conditional and marginal odds ratio could be that Z is a confounder (Fig. 1A); this would typically be the argument for adjusting for Z in the first place. Other explanations could be that Z is a mediator (Fig. 1B) or a collider (Fig. 1C). All these explanations 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 Neuhaus 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](#). 2013]

[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](#). 2008]

Regression analysis.

[[Pract Neurol](#). 2007]

Statistical hypothesis testing: associating patient characteristics with an incident condition: K [[J Wound Ostomy Continence Nurs](#)...

[See reviews...](#)

[See all...](#)

Cited by other articles in PMC

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

Is cardiovascular risk reduction therapy effective in South Asian, Chinese and other patients with diabetes? A po [[BMJ Open](#). 2017]

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](#). 2017]

A DAG-based comparison of interventional effect underestimation between composite endpoint [[BMC Medical Research Methodolo...](#)

[See all...](#)

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. Cifuentes, MD; Matthew A. Allison, MD, MPH; Rene Baudrand, MD; Joachim H. Gross, MD; 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

Younghwan Chang, MD, PhD; Seungho Ryu, MD, PhD; Yili Choi, BS; Yili Zhang, PhD; Juhee Cho, PhD; Min-Jung Kwon, MD, PhD; Young Youl Hyun, MD, PhD; Kyung-Bock Lee, MD, PhD; Hyung-Ki Kim, MD; Hyun-Suk Jung, MD; Kyung Eun Yun, MD, PhD; Jai-Hee Kim, MD, PhD; Dong-Geon Kim, PhD; Byung-Soo Cho, MD; Sung-Eun Cho, MD; Sun Cheol Chung, MD, PhD; Heecheol Choi, MD, PhD; Roberta Paolino-Santos, PhD; and Eunsoo Son, MD, PhD

Background: During 36 938 person-years of follow-up, 950 incident CKD cases were ascertained. The multivariable-adjusted differences in 5-year cumulative incidence of CKD in underweight, overweight, and obese participants compared with normal-weight participants were 1.1 (ICL -7.1 to 9.3), 3.5 (ICL 0.9 to 6.1), and 4.7 (ICL 3.0 to 10.4) cases per 1000 persons, respectively. These associations were consistently seen in all clinically relevant subgroups.

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: 62 247 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 two of the metabolic risk factors. 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 25.0 kg/m² or greater, respectively. End point outcome 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 burden and public health problem.⁽¹⁾ It is a predictor for end-stage renal disease and a risk factor for cardiovascular morbidity and mortality.⁽²⁾ Its prevalence is increasing worldwide along with the growing prevalence of obesity and metabolic disease.⁽³⁾ Individuals with metabolic syndrome, hypertension, hyperglycemia, dyslipidemia, and other metabolic abnormalities—is a major risk factor for CKD.⁽⁴⁾

Through the role of obesity-induced metabolic abnormalities in CKD development is well-established, to have metabolically healthy obese (MHO) persons, seem to have a favorable profile with no metabolic abnormalities.⁽⁵⁾ Thus, the risk for developing MHO and CKD, however, is largely unknown. The only study available found no association,⁽⁷⁾ but the comparison between MHO and non-overweight participants could be biased by selection rather than causal inference. Underweight participants, and metabolically healthy participants were defined as those with fewer than 2 metabolic risk components. Therefore, we examined the association between body mass index (BMI) and CKD in a large sample of metabolically healthy men and women who had health screening examinations.

Background: During 36 938 person-years of follow-up, 950 incident CKD cases were ascertained. The multivariable-adjusted differences in 5-year cumulative incidence of CKD in underweight, overweight, and obese participants compared with normal-weight participants were 1.1 (ICL -7.1 to 9.3), 3.5 (ICL 0.9 to 6.1), and 4.7 (ICL 3.0 to 10.4) cases per 1000 persons, respectively. These associations were consistently seen in all clinically relevant subgroups.

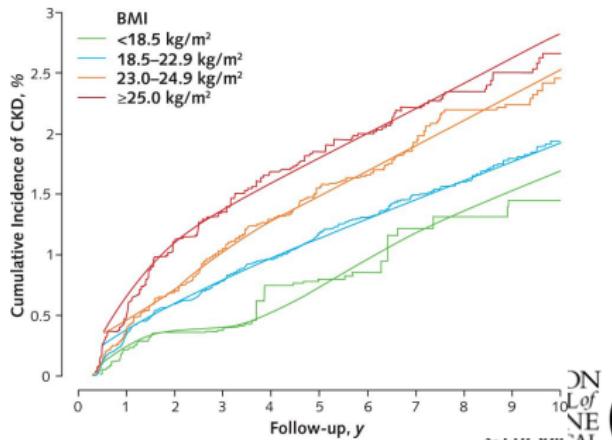
Limitation: Chronic kidney disease was identified by a single measurement at each visit.

Conclusion: Overweight and obesity are associated with an increased risk of development 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 health.

Primary Funding Source:

The Korean Samsung Health Study is a cohort study among 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 employed in various industries 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 services offered free of charge. The remaining participants registered for the screening examinations on their own.

Our analysis included all persons who had comprehensive annual or biennial health examinations between 31 December 2009 and had at least 1 other screening examination before 31 December 2013 (that is, they all had a baseline visit and a 1 follow-up visit [$n = 175 859$] in total). We excluded participants who had metabolic abnormalities ($n = 10 591$) or evidence of kidney disease at baseline ($n = 108 243$). We excluded those with fasting glucose levels of 100 mg/dL or greater or who used glucose-lowering agents. Blood pressure (BP) of



Downloaded From: <http://annals.org/> by a London School of Hygiene & Tropical Med User on 10/31/2017



Causal Inference: 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)



Causal effect

Potential Outcomes

We only observe:

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

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** or **RISK DIFFERENCE**.

Causal effect with OBSERVATIONAL data

IGNORABILITY

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

POSITIVITY

POSITIVITY: $P(A = a \mid W) > 0$ 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 $A = 0$.
- The assignment to the treatment to one unit doesn't affect the outcome of another unit (**no interference**) or **IID** random variables.
- The model used to estimate the assignment probability has to **be correctly specified**.

G-Formula, (Robins, 1986)

G-Formula for the identification of the ATE with observational data

The **ATE**=

$$\sum_w \left[\sum_y P(Y = y | A = 1, W = w) - \sum_y P(Y = y | A = 0, W = w) \right] P(W = w)$$

$$P(W = w) = \sum_{y,a} P(W = w, A = a, Y = y)$$

G-Formula

- The sums is generic notation. In reality, likely involves sums and integrals (we are just integrating out the W's).
- The **g-formula** is a **generalization of standardization** and allow to estimate unbiased treatment effect estimates.

ATE estimators

Nonparametric

- G-formula plug-in estimator (generalization of standardization).

Parametric

- Regression adjustment ([RA](#)).
- Inverse probability treatment weighting ([IPTW](#)).
- Inverse-probability treatment weighting with regression adjustment ([IPTW-RA](#)) (Kang and Schafer, 2007).

Semi-parametric Double robust (DR) methods

- Augmented inverse-probability treatment weighting (Estimation Equations) ([AIPTW](#)) (Robins, 1994).
- Targeted maximum likelihood estimation ([TMLE](#)) (**van der Laan, 2006**).

Regression-adjustment

$$\widehat{ATE}_{RA} = N^{-1} \sum_{i=1}^N [E(Y_i | A = 1, W_i) - E(Y_i | A = 0, W_i)]$$

$$m_A(w_i) = E(Y_i | A_i = A, W_i)$$

$$\widehat{ATE}_{RA} = N^{-1} \sum_{i=1}^N [\hat{m}_1(w_i) - \hat{m}_0(w_i)]$$

IPTW (Inverse probability treatment weighting)

Survey theory (Horvitz-Thompson)

$$\hat{P}_i = E(A_i | W_i) ; \text{So , } \frac{1}{\hat{p}_i} , \text{if } A = 1 \text{ and , } \frac{1}{(1 - \hat{p}_i)} , \text{if } A = 0$$

Average over the total number of individuals

$$\widehat{ATE}_{IPTW} = N^{-1} \sum_{i=1}^N \frac{A_i Y_i}{\hat{p}_i} - N^{-1} \sum_{i=1}^N \frac{(1 - A_i) Y_i}{(1 - \hat{p}_i)}$$

AIPTW (Augmented Inverse probability treatment weighting)

Solving Estimating Equations

$$\widehat{ATE}_{AIPTW} =$$

$$N^{-1} \sum_{i=1}^N [(Y(1) | A_i = 1, W_i) - (Y(0) | A_i = 0, W_i)] +$$

$$N^{-1} \sum_{i=1}^N \left(\frac{(A_i = 1)}{P(A_i = 1 | W_i)} - \frac{(A_i = 0)}{P(A_i = 0 | W_i)} \right) [Y_i - E(Y | A_i, W_i)]$$



Nonparametric

- Course of dimensionality (sparsity: zero empty cell)

Parametric

- Parametric models are misspecified (all models are wrong but some are useful, Box, 1976), and break down for high-dimensional data.
- (RA) Issue: extrapolation and biased if misspecification, no information about treatment mechanism.
- (IPTW) Issue: sensitive to course of dimensionality, inefficient in case of extreme weights and biased if misspecification. Non information about the outcome.

Double-robust (DR) estimators

Prons: Semi-parametric Double-Robust Methods

- DR methods give **two chances at consistency** if any of two nuisance parameters is consistently estimated.
- DR methods are **less sensitive** to course of dimensionality.

Cons: Semi-parametric Double-Robust Methods

- DR methods are unstable and inefficient if the propensity score (PS) is small (**violation of positivity assumption**) (vand der Laan, 2007).
- AIPTW and IPTW-RA do not respect the **limits of the boundary space of Y**.
- **Poor performance if dual misspecification** (Benkeser, 2016).

Targeted Maximum Likelihood Estimation (TMLE)

Pros: TMLE

- (TMLE) is a general algorithm for the construction of **double-robust**, **semiparametric** MLE, efficient **substitution** estimator (Van der Laan, 2011)
- Better performance than competitors has been largely documented (Porter, et. al., 2011).
- (TMLE) **Respects bounds on Y**, less sensitive to **misspecification** and to **near-positivity** violations (Benkeser, 2016).
- (TMLE) **Reduces bias** through **ensemble learning** if misspecification, even dual misspecification.
- For the ATE, **Inference** is based on the **Efficient Influence Curve**. Hence, the **CLT** applies, making inference easier.

Cons: TMLE

- The procedure is only available in R: **tmle** package (Gruber, 2011).

Targeted learning

Springer Series in Statistics

Targeted Learning

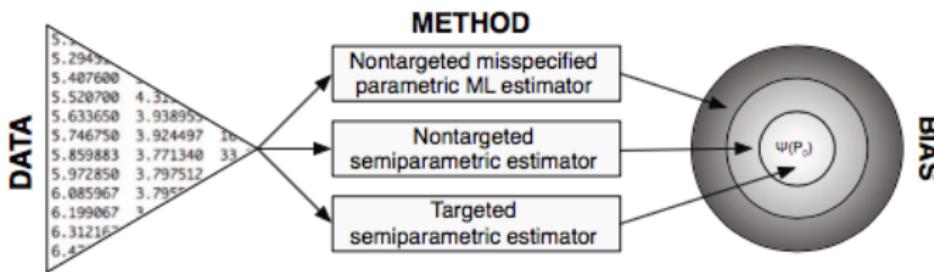
Causal Inference for Observational
and Experimental Data

 Springer

Source: Mark van der Laan and Sherri Rose. Targeted learning: causal inference for observational and experimental data. Springer Series in Statistics, 2011.



Why Targeted learning?



Source: Mark van der Laan and Sherri Rose. Targeted learning: causal inference for observational and experimental data. Springer Series in Statistics, 2011.

TMLE ROAD MAP

MC simulations: Luque-Fernandez et al, 2017 (in press, American Journal of Epidemiology)

	ATE		BIAS (%)		RMSE		95%CI coverage (%)	
	N=1,000	N=10,000	N=1,000	N=10,000	N=1,000	N=10,000	N=1,000	N=10,000
First scenario* (correctly specified models)								
True ATE	-0.1813							
Naïve	-0.2234	-0.2218	23.2	22.3	0.0575	0.0423	77	89
AIPYW	-0.1843	-0.1848	1.6	1.9	0.0534	0.0180	93	94
IPTW-RA	-0.1831	-0.1838	1.0	1.4	0.0500	0.0174	91	95
TMLE	-0.1832	-0.1821	1.0	0.4	0.0482	0.0158	95	95
Second scenario ** (misspecified models)								
True ATE	-0.1172							
Naïve	-0.0127	-0.0121	89.2	89.7	0.1470	0.1100	0	0
BFit AIPYW	-0.1155	-0.0920	1.5	11.7	0.0928	0.0773	65	65
BFit IPTW-RA	-0.1268	-0.1192	8.2	1.7	0.0442	0.0305	52	73
TMLE	-0.1181	-0.1177	0.8	0.4	0.0281	0.0107	93	95

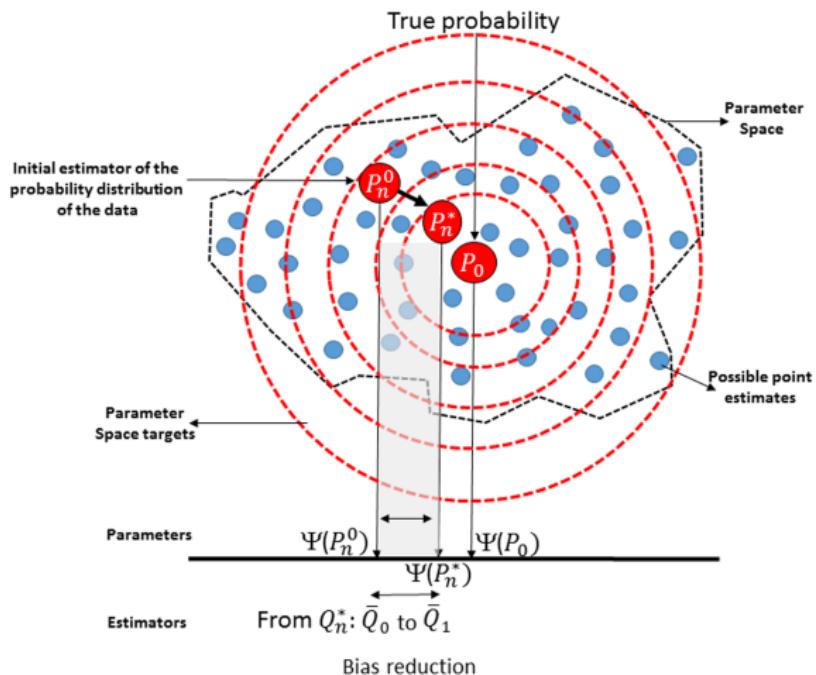
*First scenario : correctly specified models and near-positivity violation

**Second scenario: misspecification, near-positivity violation and adaptive model selection

Source: 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>



TMLE ROAD MAP



Substitution estimation: $\hat{E}(Y | A, W)$

- First compute the outcome regression $E(Y | A, W)$ using the **Super-Learner** to then derive the Potential Outcomes and compute $\Psi^{(0)} = E(Y(1) | A = 1, W) - E(Y(0) | A = 0, W)$.
- Estimate the exposure mechanism $P(A=1|W)$ using the **Super-Learner** to predict the values of the propensity score.
- Compute $HAW = \left(\frac{\mathbb{I}(A_i=1)}{P(A_i=1|W_i)} - \frac{\mathbb{I}(A_i=0)}{P(A_i=0|W_i)} \right)$ for each individual, named the **clever covariate H**.



Fluctuation step: Epsilon

Fluctuation step ($\hat{\epsilon}_0$, $\hat{\epsilon}_1$)

- Update $\Psi^{(0)}$ through a fluctuation step incorporating the information from the exposure mechanism:

$$\mathbf{H}(1)\mathbf{W} = \frac{\mathbb{I}(A_i=1)}{\hat{P}(A_i=1|W_i)} \text{ and, } \mathbf{H}(0)\mathbf{W} = -\frac{\mathbb{I}(A_i=0)}{\hat{P}(A_i=0|W_i)}.$$

- This step aims to **reduce bias** minimising the mean squared error (MSE) for (Ψ) and considering the **bounds of the limits of Y**.
- The fluctuation parameters ($\hat{\epsilon}_0$, $\hat{\epsilon}_1$) are estimated using maximum likelihood procedures (in Stata):

```
. glm Y HAW, fam(binomial) nocons offset(E(Y| A, W))  
. mat e = e(b),  
. gen double ε = e[1, 1],
```

Targeted estimate of the ATE ($\hat{\Psi}$)

$\Psi^{(0)}$ update using ϵ (epsilon)

$$\mathbf{E}^*(Y | A = 1, W) = \text{expit} [\text{logit}[E(Y | A = 1, W)] + \hat{\epsilon}_1 H_1(1, W)]$$

$$\mathbf{E}^*(Y | A = 0, W) = \text{expit} [\text{logit}[E(Y | A = 0, W)] + \hat{\epsilon}_0 H_0(0, W)]$$

Targeted estimate of the ATE from $\Psi^{(0)}$ to $\Psi^{(1)}$: ($\hat{\Psi}$)

$$\Psi^{(1)} : \hat{\Psi} = [\mathbf{E}^*(Y(1) | A = 1, W) - \mathbf{E}^*(Y(0) | A = 0, W)]$$

TMLE inference: INFLUENCE CURVE

M-ESTIMATORS: Semi-parametric and Empirical processes theory

An estimator is **asymptotically linear** with **influence function φ (IC)** if the estimator can be **approximate by an empirical average** in the sense that

$$(\hat{\theta} - \theta_0) = \frac{1}{n} \sum_{i=1}^n (\text{IC}) + O_p(1/\sqrt{n})$$

(Bickel, 1997).

TMLE inference: Bickel (1993); Tsiatis (2007); Van der Laan (2011); Kennedy (2016)

- The **IC** estimation is a more general approach than M-estimation.
- The **Efficient IC** has mean zero $E(\text{IC}_{\hat{\psi}}(y_i, \psi_0)) = 0$ and **finite variance**.
- By the **Weak Law of the Large Numbers**, the **Op** converges to zero in a rate $1/\sqrt{n}$ as $n \rightarrow \infty$ (Bickel, 1993).
- The **Efficient IC** requires **asymptotically linear** estimators.

TMLE inference: Influence curve

TMLE inference

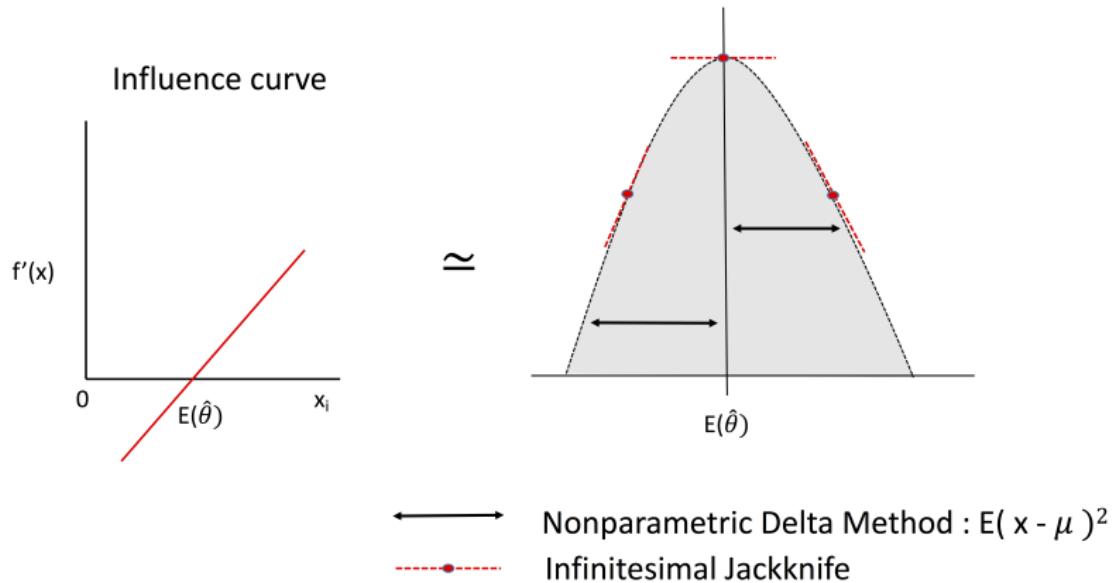
$$\text{IC} = \left(\frac{(A_i = 1)}{P(A_i = 1 | W_i)} - \frac{(A_i = 0)}{P(A_i = 0 | W_i)} \right) [Y_i - E_1(Y | A_i, W_i)] + \\ [E_1(Y(1) | A_i = 1, W_i) - E_1(Y(0) | A_i = 0, W_i)] - \psi$$

$$\text{Standard Error : } \sigma(\psi_0) = \frac{SD(IC_n)}{\sqrt{n}}$$

TMLE inference

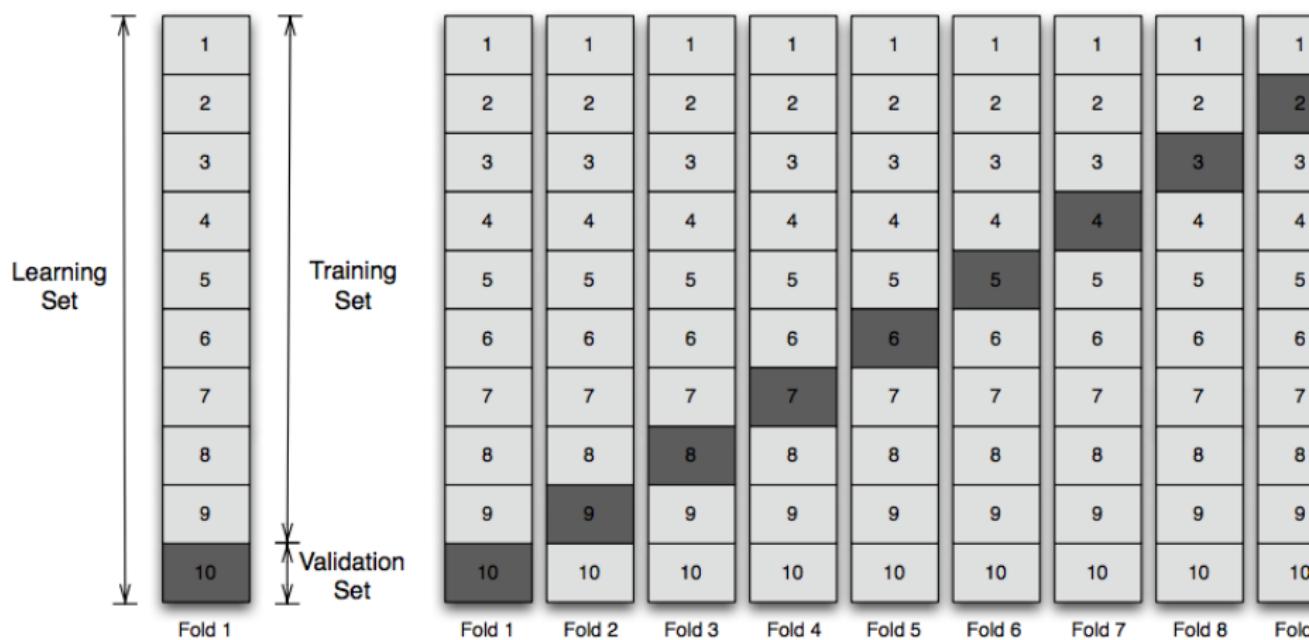
- The **Efficient IC**, first introduced by Hampel (1974), is used to apply readily the **CLT** for statistical inference using TMLE.
- The **Efficient IC** is the same as the infinitesimal jackknife and the **nonparametric delta method**. Also named the "**canonical gradient**" of the pathwise derivative of the target parameter ψ or "**approximation by averages**"(Efron, 1982).

IC: Geometric interpretation



Estimate of the ψ Standard Error using the efficient Influence Curve.
Image credit: Miguel Angel Luque-Fernandez

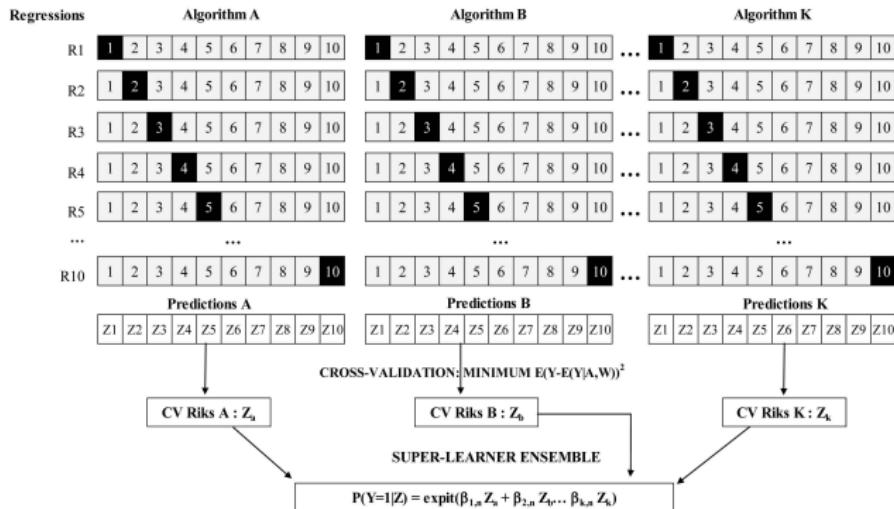
Targeted learning



Source: Mark van der Laan and Sherri Rose. Targeted learning: causal inference for observational and experimental data. Springer Series in Statistics, 2011.



Super-Learner: Ensemble learning



To apply the EIC we need data-adaptive estimation for both, the model of the outcome, and the model of the treatment.

Asymptotically, the final weighted combination of algorithms (Super Learner) performs as well as or better than the best-fitting algorithm (van der Laan, 2007).

Luque-Fernandez, MA. 2017. TMLE steps adapted from Van der Laa, 2011.



Ensemble Learning Targeted Maximum Likelihood Estimation

- **eltmle** is a Stata program implementing R-TMLE for the ATE for a binary or continuous outcome and binary treatment.
- **eltmle** includes the use of a **super-learner**(Polley E., et al. 2011).
- I used the default Super-Learner algorithms implemented in the base installation of the tmle-R package v.1.2.0-5 (Susan G. and Van der Laan M., 2007).
- i) stepwise selection, ii) GLM, iii) a GLM interaction.
- Additionally, **eltmle** users will have the option to include Bayes GLM and GAM.



Stata Implementation: overall structure

```
45
46 capture program drop eltmle
47 program define eltmle
48     syntax [varlist] [if] [pw] [, slaipw slaipwbgam tmle tmlebgam]
49     version 13.2
50     marksample touse
51     local var `varlist' if `touse'
52     tokenize `var'
53     local yvar = "`1'"
54     global flag = cond(`yvar'<=1,1,0)
55     qui sum `yvar'
56     global b = `r(max)'
57     global a = `r(min)'
58     qui replace `yvar' = (`yvar' - `r(min)') / (`r(max)' - `r(min)') if `yvar'>1
59     local dir `c(pwd)'
60     cd `dir'
61     qui export delimited `var' using "data.csv", nolabel replace
62     if "`slaipw'" == "" & "`slaipwbgam'" == "" & "`tmlebgam'" == "" {
63         tmle `varlist'
64     }
65     else if "`tmlebgam'" == "tmlebgam" {
66         tmlebgam `varlist'
67     }
68     else if "`slaipw'" == "slaipw" {
69         slaipw `varlist'
70     }
71     else if "`slaipwbgam'" == "slaipwbgam" {
72         slaipwbgam `varlist'
73     }
74 end
```

Stata Implementation: calling the SL

```
program tmle
// Write R Code dependencies: foreign Surperlearner
set more off
qui: file close _all
qui: file open rcode using SLS.R, write replace
qui: file write rcode ///
    `"set.seed(123)"' _newline ///
    `"list.of.packages <- c("foreign","SuperLearner")"' _newline ///
    `"new.packages <- list.of.packages[!(list.of.packages %in% installed.packages() [, "Package"])]"' _newline ///
    `"if(length(new.packages)) install.packages(new.packages, repos='http://cran.us.r-project.org')"' _newline ///
    `"library(SuperLearner)"' _newline ///
    `"library(foreign)"' _newline ///
    `"data <- read.csv("data.csv", sep=",")"' _newline ///
    `"attach(data)"' _newline ///
    `"SL.library <- c("SL.glm","SL.step","SL.glm.interaction")"' _newline ///
    `"n <- nrow(data)"' _newline ///
    `"nvar <- dim(data)[2]"' _newline ///
    `"Y <- data[,1]"' _newline ///
    `"A <- data[,2]"' _newline ///
    `"X <- data[,2:nvar]"' _newline ///
    `"W <- data[,3:nvar]"' _newline ///
    `"X1 <- X0 <- X"' _newline ///
    `"X1[,1] <- 1"' _newline ///
    `"X0[,1] <- 0"' _newline ///
    `"newdata <- rbind(X,X1,X0)"' _newline ///
    `"Q <- try(SuperLearner(Y = data[,1], X = X, SL.library=SL.library, family=binomial(), newX=newdata, method="method2"))"' _newline ///
    `"Q <- as.data.frame(Q[[4]])"' _newline ///
    `"QAW <- Q[1:n,]"' _newline ///
    `"QIW <- Q[((n+1):(2*n)),]"' _newline ///
    `"QOW <- Q[((2*n+1):(3*n)),]"' _newline ///
    `"g <- suppressWarnings(SuperLearner(Y = data[,2], X = W, SL.library = SL.library, family = binomial(), method = "method2"))"' _newline ///
    `"ps <- g[[4]]"' _newline ///
    `"ps[ps<0.025] <- 0.025"' _newline ///
    `"ps[ps>0.975] <- 0.975"' _newline ///
    `"data <- cbind(data,QAW,QIW,QOW,ps,Y,A)"' _newline ///
    `"write.dta(data, "data2.dta")"' _newline
qui: file close rcode
```

Stata Implementation: Batch file executing R

```
112 qui: file close rcode
113
114 // Write batch file to find R.exe path and R version
115 set more off
116 qui: file close _all
117 qui: file open bat using setup.bat, write replace
118 qui: file write bat ///
119 `"@echo off" _newline ///
120 `SET PATHROOT=C:\Program Files\R\``_newline ///
121 `echo Locating path of R...``_newline ///
122 `echo.``_newline ///
123 `if not exist "%PATHROOT%" goto:NO_R``_newline ///
124 `for /f "delims=%" %%r in (' dir /b "%PATHROOT%R*" ') do ("`_newline ///
125 `echo Found %%r``_newline ///
126 `echo shell "%PATHROOT%&rlbin\x64\R.exe" CMD BATCH SLS.R > runr.do``_newline ///
127 `echo All set!``_newline ///
128 `goto:DONE``_newline ///
129 `)``_newline ///
130 `:NO_R``_newline ///
131 `echo R is not installed in your system.`_newline ///
132 `echo.``_newline ///
133 `echo Download it from https://cran.r-project.org/bin/windows/base/``_newline ///
134 `echo Install it and re-run this script``_newline ///
135 `:DONE``_newline ///
136 `echo.``_newline ///
137 `pause``
138 qui: file close bat
139
140 //Run batch
141 shell setup.bat
142 //Run R
143 do runr.do
144
145 // Read Revised Data Back to Stata
146 clear
147 quietly: use "data2.dta", clear
148
149 // Q to logit scale
150 gen logQAW = log(QAW / (1 - QAW))
151 gen logQ1W = log(Q1W / (1 - Q1W))
152 gen logQ0W = log(Q0W / (1 - Q0W))
153
154 // Clever covariate HAW
```

Syntax eltmle Stata command

eltmle Y A W [, slapiw slapwbgam tmle tmlebgam]

Y: Outcome: numeric binary or continuous variable.

A: Treatment or exposure: numeric binary variable.

W: Covariates: vector of numeric and categorical variables.



Output for continuous outcome

```
.use http://www.stata-press.com/data/r14/cattaneo2.dta  
.eltmle bweight mbsmoke mage medu prenatal mmarrried, tmle
```

Variable	Obs	Mean	Std. Dev.	Min	Max
-----+-----					
POM1	4,642	2832.384	74.56757	2580.186	2957.627
POM0	4,642	3063.015	89.53935	2868.071	3167.264
WT	4,642	-.0409955	2.830591	-6.644464	21.43709
PS	4,642	.1861267	.110755	.0372202	.8494988

ACE:

Additive Effect: -230.63; Estimated Variance: 600.93; p-value: 0.0000;
95%CI: (-278.68, -182.58)

Risk Differences:-0.0447; SE: 0.0047; p-value: 0.0000;
95%CI:(-0.05, -0.04)



Simulations comparing Stata ELTMLE vs R-TMLE

```
. mean psi aipw slaipw tmle
Mean estimation
Number of obs      = 1,000
-----
|   Mean
+-----+
    True |   .173
    aipw |   .170
    slaipw |   .170
  Stata-tmle |   .170
-----
R-TMLE |   .170
-----
```



ONLINE open free tutorial

Link to the tutorial

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

Stata Implementation: source code

<https://github.com/migariane/meltmle> for MAC users

<https://github.com/migariane/weltmle> for Windows users

Stata installation and step by step commented syntax

github install migariane/meltmle (For MAC users)

github install migariane/weltmle (For Windows users)

which eltmle

viewsource eltmle.ado



One sample simulation: TMLE reduces bias

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



References

References

- ① Bickel, Peter J.; Klaassen, Chris A.J.; Ritov, Yaacov; Wellner Jon A. (1997). Efficient and adaptive estimation for semiparametric models. New York: Springer.
- ② Hamble, F.R., (1974). The influence curve and its role in robust estimation. *J Amer Statist Asso.* 69, 375-391.
- ③ Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Amer Statist Assoc.* 1994;89:846866.
- ④ Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics.* 2005;61:962972.
- ⑤ Tsiatis AA. Semiparametric Theory and Missing Data. Springer; New York: 2006
- ⑥ Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science.* 2007;22(4):523539
- ⑦ Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology.* 1974;66:688701



References

References

- ① Luque-Fernandez, Miguel Angel. (2017). Targeted Maximum Likelihood Estimation for a Binary Outcome: Tutorial and Guided Implementation.
- ② StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.
- ③ Gruber S, Laan M van der. (2011). Tmle: An R package for targeted maximum likelihood estimation. UC Berkeley Division of Biostatistics Working Paper Series.
- ④ Laan M van der, Rose S. (2011). Targeted learning: Causal inference for observational and experimental data. Springer Series in Statistics.626p.
- ⑤ Van der Laan MJ, Polley EC, Hubbard AE. (2007). Super learner. Statistical applications in genetics and molecular biology 6.
- ⑥ Bickel, Peter J.; Klaassen, Chris A.J.; Ritov, Yaacov; Wellner Jon A. (1997). Efficient and adaptive estimation for semiparametric models. New York: Springer.
- ⑦ E. H. Kennedy. Semiparametric theory and empirical processes in causal inference. In: Statistical Causal Inferences and Their Applications in Public Health Research, in press.



Thank YOU

