Big Data, Data Science, and Causal Inference: A Primer for Clinicians

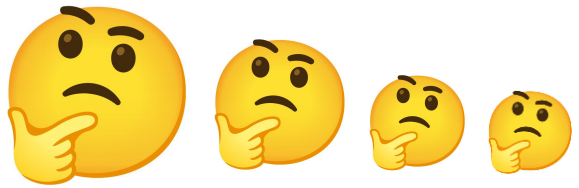
Yoshihiko Raita [1], Carlos A. Camargo Jr. [1,2,3], Liming Liang [1,3,4] and Kohei Hasegawa [1,3,4]*

[1] Department of Emergency Medicine, Harvard Medical School, Massachusetts General Hospital, Boston, MA, United States, [2] Division of Rheumatology, Allergy, and Immunology, Department of Medicine, Harvard Medical School, Massachusetts General Hospital, Boston, MA, United States, [3] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, United States, [4] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States

Nat Sirirutbunkajorn
Radiation oncologist, Ramathibodi hospital
Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol university
Nut19012537@gmail.com

Background

- **Big data** is emerging as the next thing to transform medicine into precision medicine.
- **Precision medicine** using big data **cannot be achieved** by algorithms that operate exclusively in data-driven prediction modes, as do most machine learning algorithms.
  - Why though?

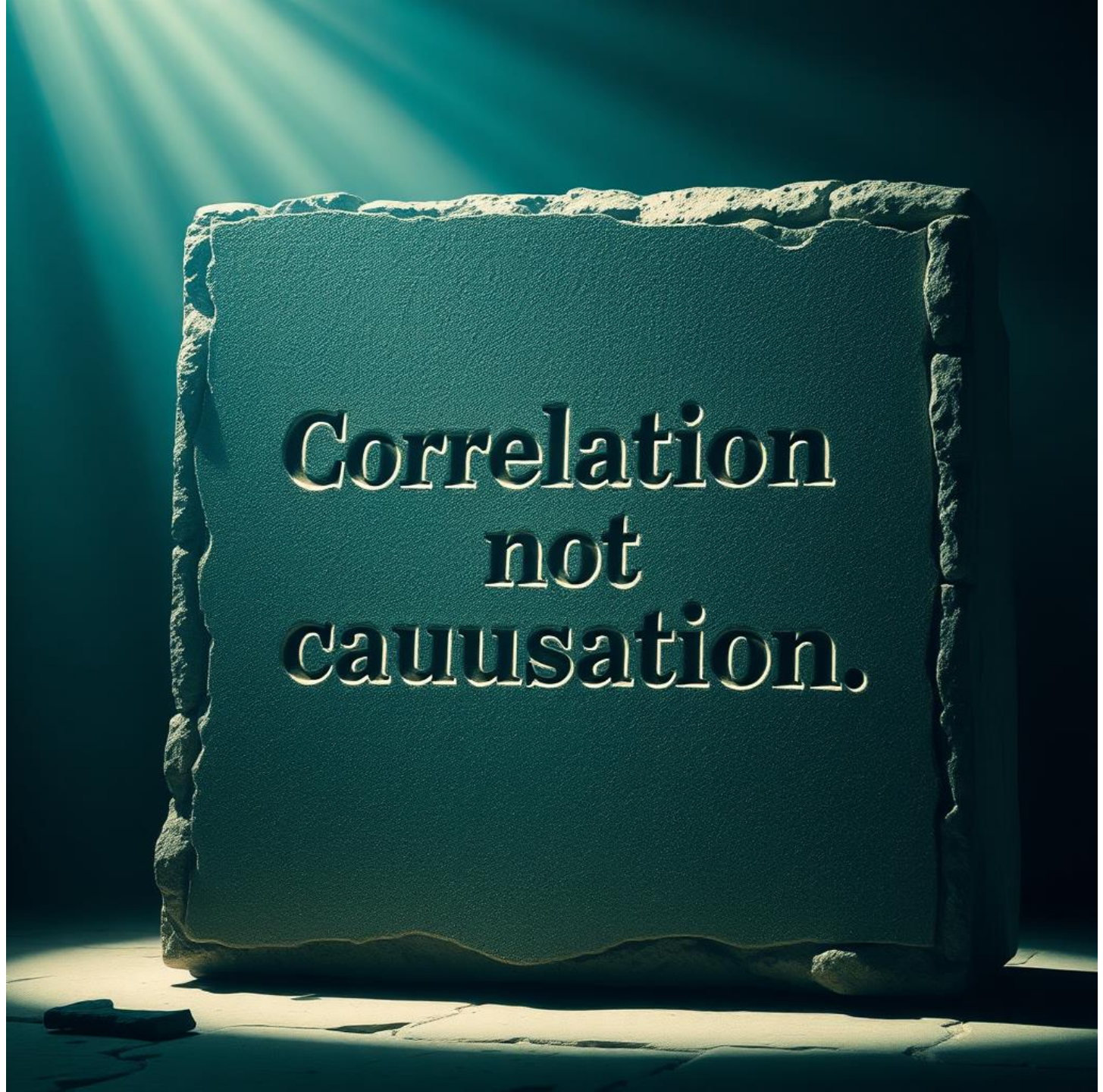


Some history

1920s

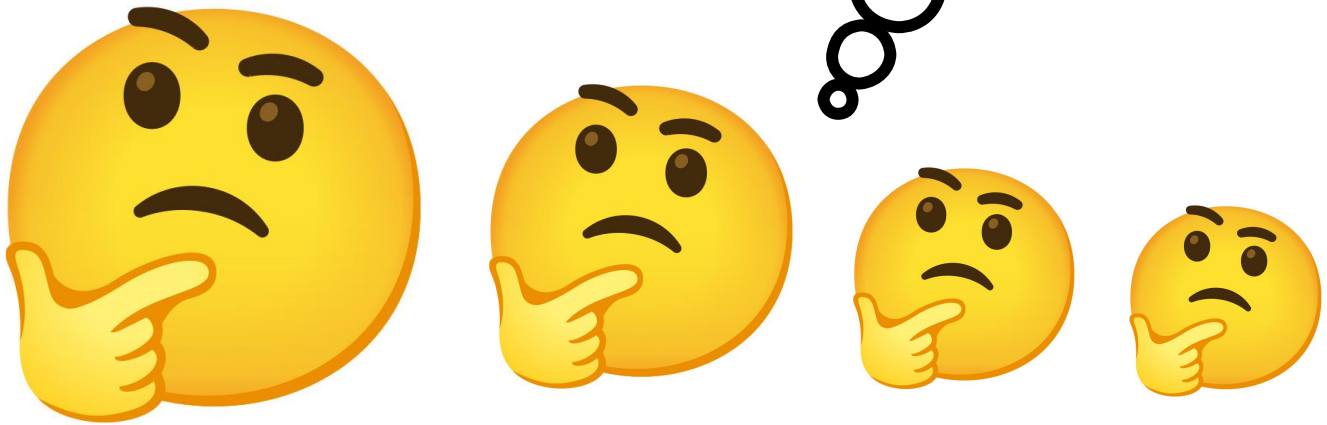
RCT's can be used to study cause and effect but not observational data

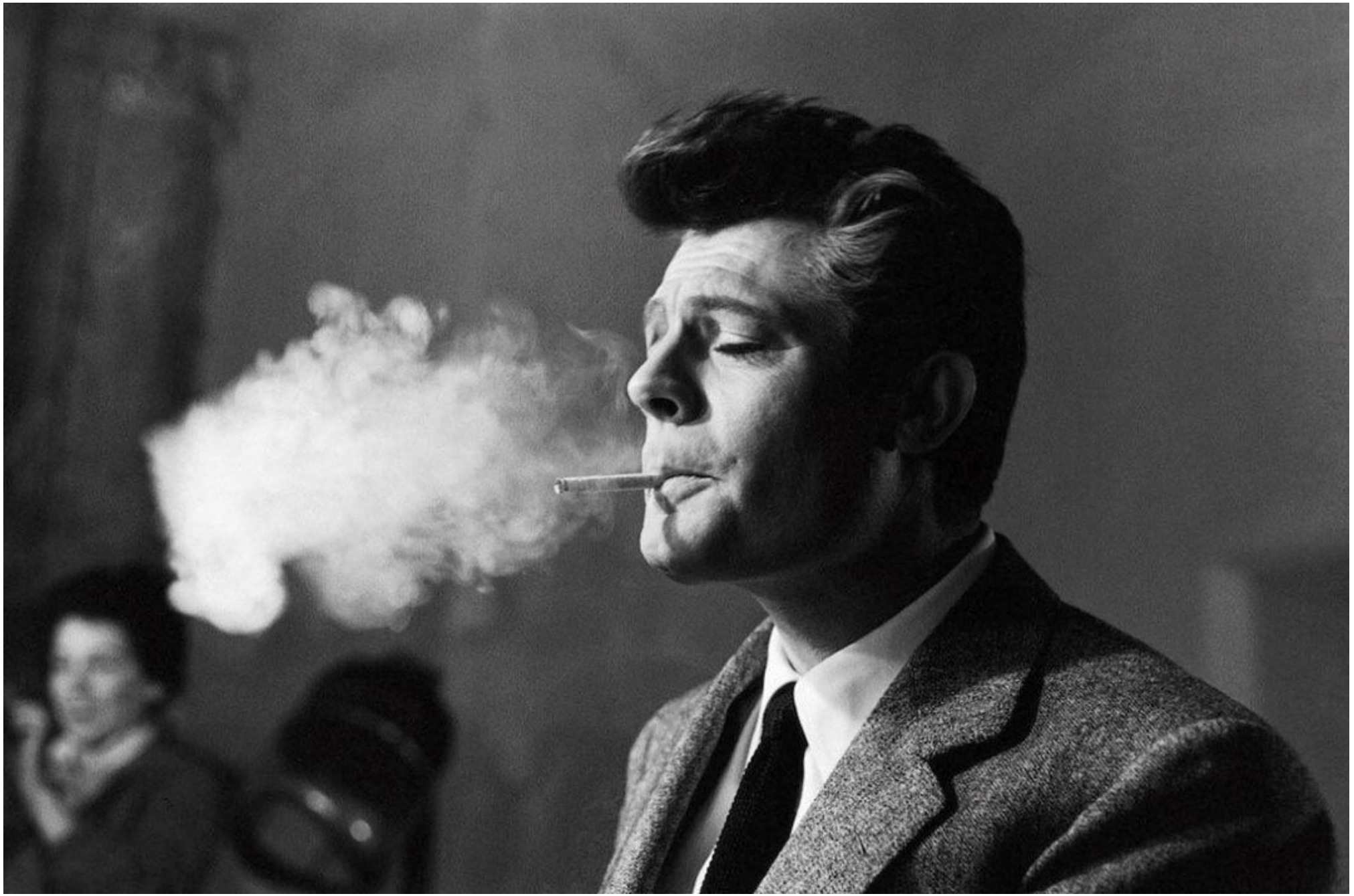
Ronald Fisher, father of statistics

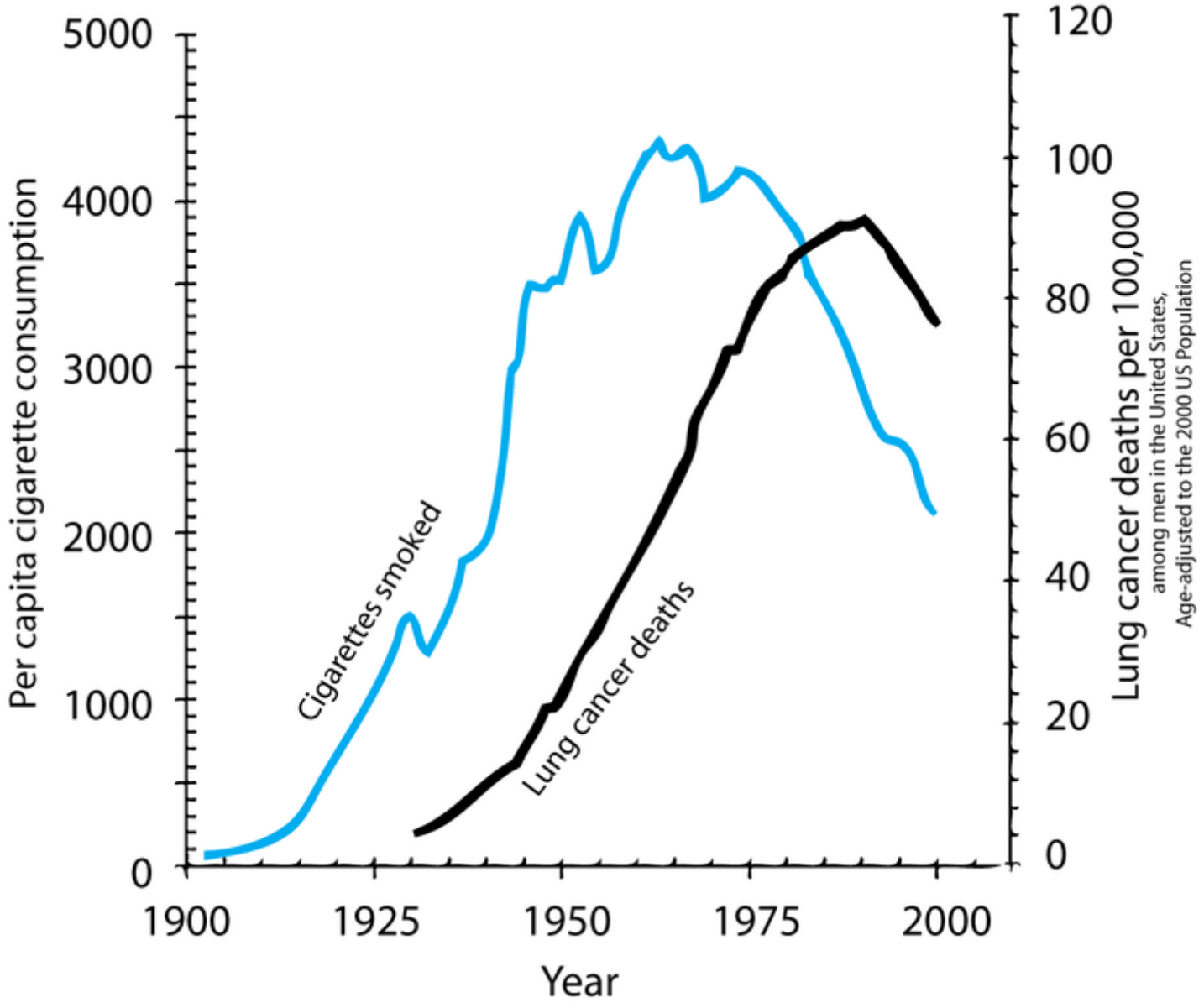

Correlation
not
cauusation.

Background

When can we say that a thing is a cause and not just an association?






5000
4000
3000
2000
1000
0
Per capita cigarette consumption
120
100
80
60
40
20
0
Lung cancer deaths per 100,000
among men in the United States,
Age-adjusted to the 2000 US Population
Cigarettes smoked
Lung cancer deaths
1900
1925
1950
1975
2000
Year

BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

Discussion

To summarize, it is not reasonable, in our view, to attribute the results to any special selection of cases or to bias in recording. In other words, it must be concluded that there is a real association between carcinoma of the lung and smoking. Further, the comparison of the smoking habits of patients in different disease groups, shown in Table X, revealed no association between smoking and other respiratory diseases or between smoking and cancer of the other sites (mainly stomach and large bowel). The

https://www.nytimes.com/2008/10/07/business/media/07adco.html

When Genius Errs: R. A. Fisher and the Lung Cancer Controversy

FISHER'S ARGUMENTS CONCERNING LUNG CANCER

Fisher developed four lines of argument in questioning the causal relation of lung cancer to smoking. I will first list these and then briefly describe the evidence he produced in support of these arguments.

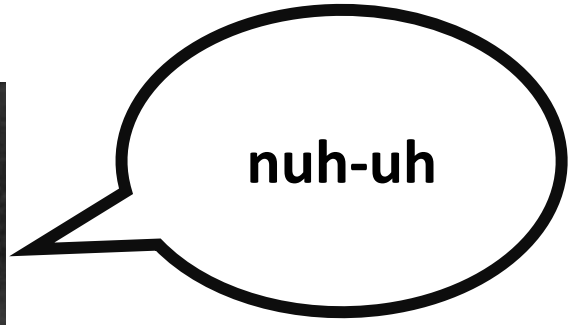
1) If **A** is associated with **B**, then not only is it possible that **A** causes **B**, but it is also possible that **B** is the cause of **A**. In other words, smoking may cause lung cancer, but it is a logical possibility that lung cancer causes smoking.

2) There may be a genetic predisposition to smoke (and that genetic predisposition is presumably also linked to lung cancer).

3) Smoking is unlikely to cause lung cancer because secular trend and other ecologic data do not support this relation.

4) Smoking does not cause lung cancer because inhalers are less likely to develop lung cancer than are noninhalers (9).

Fisher sees the argument that lung cancer causes smoking as an essentially unsupported speculation. His view is best described in his own words:

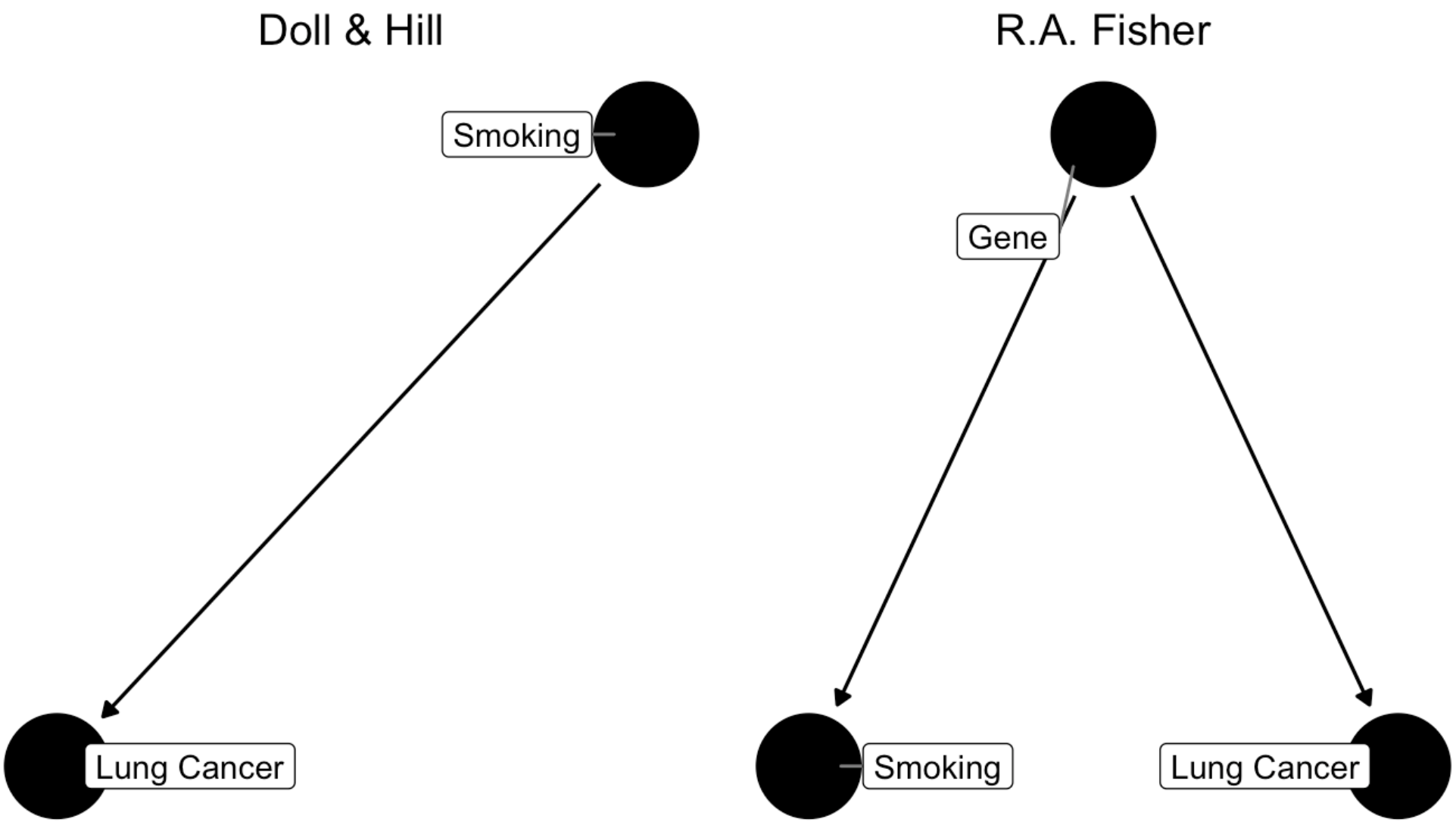


https://pubmed.ncbi.nlm.nih.gov/2000852/

When Genius Errs: R. A. Fisher and the Lung Cancer Controversy

This was written by the same man who tried to bully Neyman! When his first letter to the *British Medical Journal* was attacked and he was impugned for taking a fee from the tobacco industry, it probably fixed his views. His daughter mentions how offended he was by the rebuttal letters that pointed out he was a paid consultant.

Secondly, Fisher was a political conservative and an elitist (as were most eugenicists) and was disturbed by the British Medical Association's appeal to censure cigarette advertising and launch a public health campaign against smoking (Fisher was a smoker of pipes and cigarettes). He compares this proposed public education campaign to totalitarian propaganda and complains it is premature in a letter to the *British Medical Journal*:


Doll & Hill
Smoking
Lung Cancer
R.A. Fisher
Gene
Smoking
Lung Cancer

In 1964, the US surgeon general release a report on the effect of smoking and health

SMOKING *and* HEALTH

REPORT OF THE ADVISORY COMMITTEE
TO THE SURGEON GENERAL
OF THE PUBLIC HEALTH SERVICE

U.S DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

https://www.unav.edu/documents/16089811/16155256/Smoking+and+Health+the+Surgeon+General+Report+1964.pdf

Statistical methods cannot establish proof of a causal relationship in an association. The causal significance of an association is a matter of judgment which goes beyond any statement of statistical probability. To judge or evaluate the causal significance of the association between the attribute or agent and the disease, or effect upon health, a number of criteria must be utilized, no one of which is an all-sufficient basis for judgment. These criteria include:

a) The consistency of the association
b) The strength of the association
c) The specificity of the association
d) The temporal relationship of the association
e) The coherence of the association

These criteria were utilized in various sections of this Report. The most extensive and illuminating account of their utilization is to be found in Chapter 9 in the section entitled "Evaluation of the Association Between Smoking and Lung Cancer".

https://www.unav.edu/documents/16089811/16155256/Smoking+and+Health+the+Surgeon+General+Report+1964.pdf

Statistical methods cannot establish proof of a causal relationship in an association. The causal significance of an association is a matter of judgment which goes beyond any statement of statistical probability. To judge or evaluate the causal significance of the association between the attribute or agent and the disease, or effect upon health, a number of criteria must be utilized, no one of which is an all-sufficient basis for judgment. These criteria include:

a) The consistency of the association
b) The strength of the association
c) The specificity of the association
d) The temporal relationship of the association
e) The coherence of the association

These criteria were utilized in various sections of this Report. The most extensive and illuminating account of their utilization is to be found in Chapter 9 in the section entitled "Evaluation of the Association Between Smoking and Lung Cancer".

Bradford Hill criteria
1. Strength
2. Consistency
3. Specificity
4. Temporality
5. Biological gradient
6. Plausibility
7. Coherence
8. Experiment
9. Analogy

https://www.unav.edu/documents/16089811/16155256/Smoking+and+Health+the+Surgeon+General+Report+1964.pdf

Lung Cancer

Cigarette smoking is causally related to lung cancer in men; the magnitude of the effect of cigarette smoking far outweighs all other factors. The data for women, though less extensive, point in the same direction.

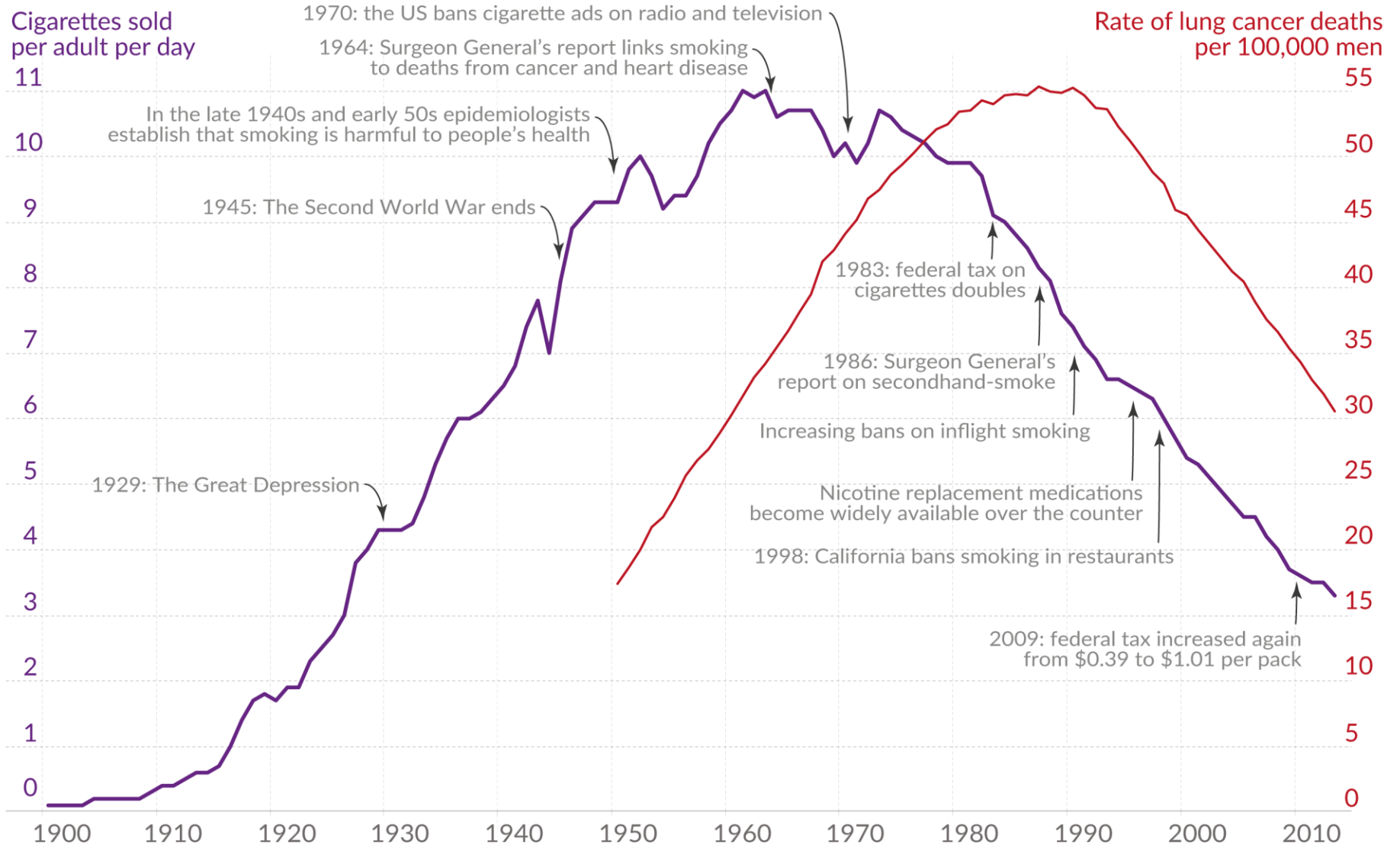
The Committee's Judgment in Brief

On the basis of prolonged study and evaluation of many lines of converging evidence, the Committee makes the following judgment:

Cigarette smoking is a health hazard of sufficient importance in the United States to warrant appropriate remedial action.

Cigarette sales and lung cancer mortality in the US

Our World in Data



Data sources: International Smoking Statistics (2017); WHO Cancer Mortality Database (IARC). The death rate from lung-cancer is age-standardized.

OurWorldinData.org – Research and data to make progress against the world's largest problems.

https://csslab.uc3m.es/dataviz/projects/2023/100407614/

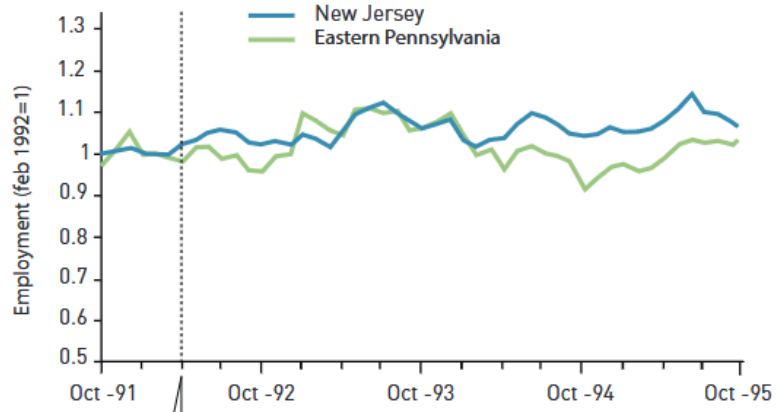
Nobel Prize in economics explodes minimum wage and jobs myth

The award of this year's Nobel Prize in economics has further exploded a decades-old myth that increasing minimum wages costs jobs.

The effect of increasing the minimum wage

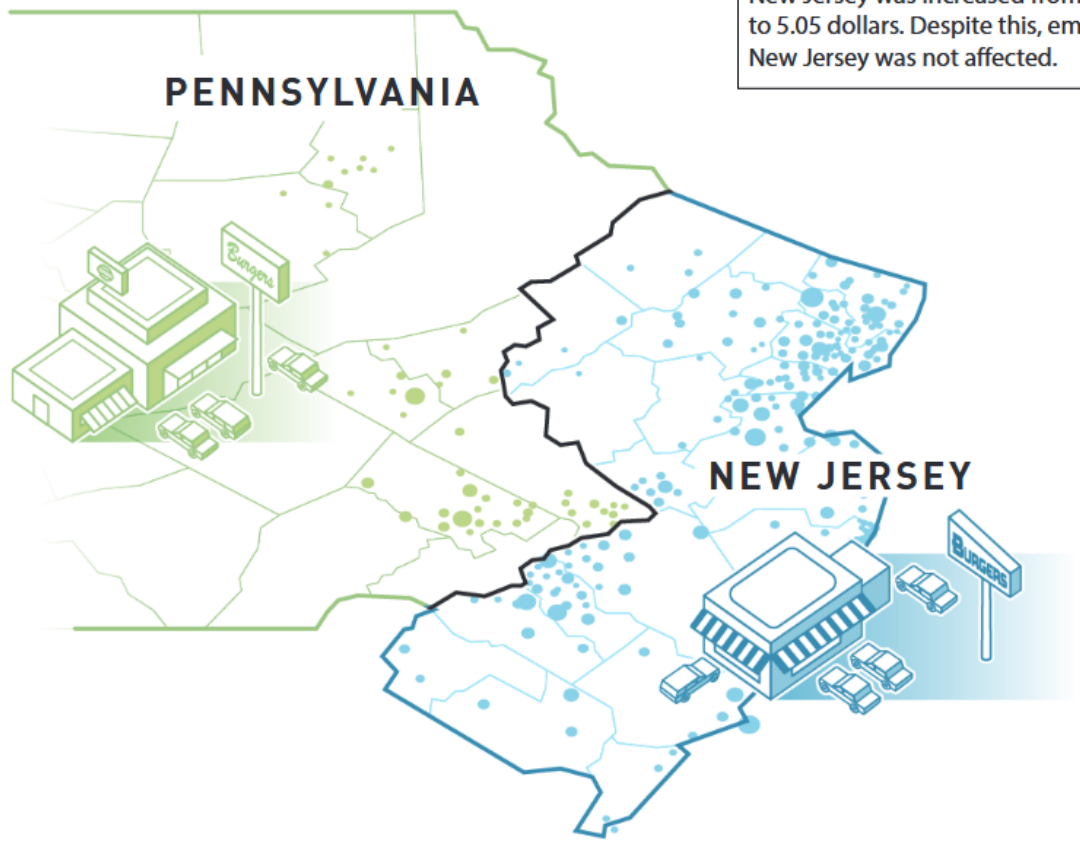
Card and Krueger used a natural experiment to study how increasing the minimum wage affects employment.

The researchers identified a treatment group (restaurants in New Jersey) and a control group (restaurants in eastern Pennsylvania) to measure the effect of increasing the minimum wage.



1 April 1992: The hourly minimum wage in New Jersey was increased from 4.25 dollars to 5.05 dollars. Despite this, employment in New Jersey was not affected.

CONTROL GROUP TREATMENT GROUP



Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania

By DAVID CARD AND ALAN B. KRUEGER*

On April 1, 1992, New Jersey's minimum wage rose from $4.25 to $5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above $5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment. (*JEL* J30, J23)

III. Employment Effects of the Minimum-Wage Increase

A. *Differences in Differences*

Table 3 summarizes the levels and changes in average employment per store in

B. *Regression-Adjusted Models*

The comparisons in Table 3 make no allowance for other sources of variation in employment growth, such as differences across chains. These are incorporated in the estimates in Table 4. The entries in this table are regression coefficients from mod-

The Effects of Naloxone Access Laws on Opioid Abuse, Mortality, and Crime*

Jennifer L. Doleac Anita Mukherjee

78 pages

August 12, 2021

The U.S. is experiencing an epidemic of opioid abuse. In response, states have implemented a variety of policies including increased access to naloxone, a drug that can save lives when administered during an overdose. There is a concern that widespread naloxone access may unintentionally lead to increased or riskier opioid use by reducing the risk of death from overdose, however. In this paper, we use the staggered timing of state-level naloxone access laws as a natural experiment to measure the effects of broadening access to this lifesaving drug. We find that broadened access led to more opioid-related emergency room visits and more opioid-related theft, with no net measurable reduction in opioid-related mortality. We conclude that naloxone has a clear and important role in harm-reduction, yet its ability to combat the opioid epidemic's death toll may be limited without complementary efforts.

JEL Codes: I18, K42, D81

Ryan Marino 💊🔪 @RyanMarino · Mar 6

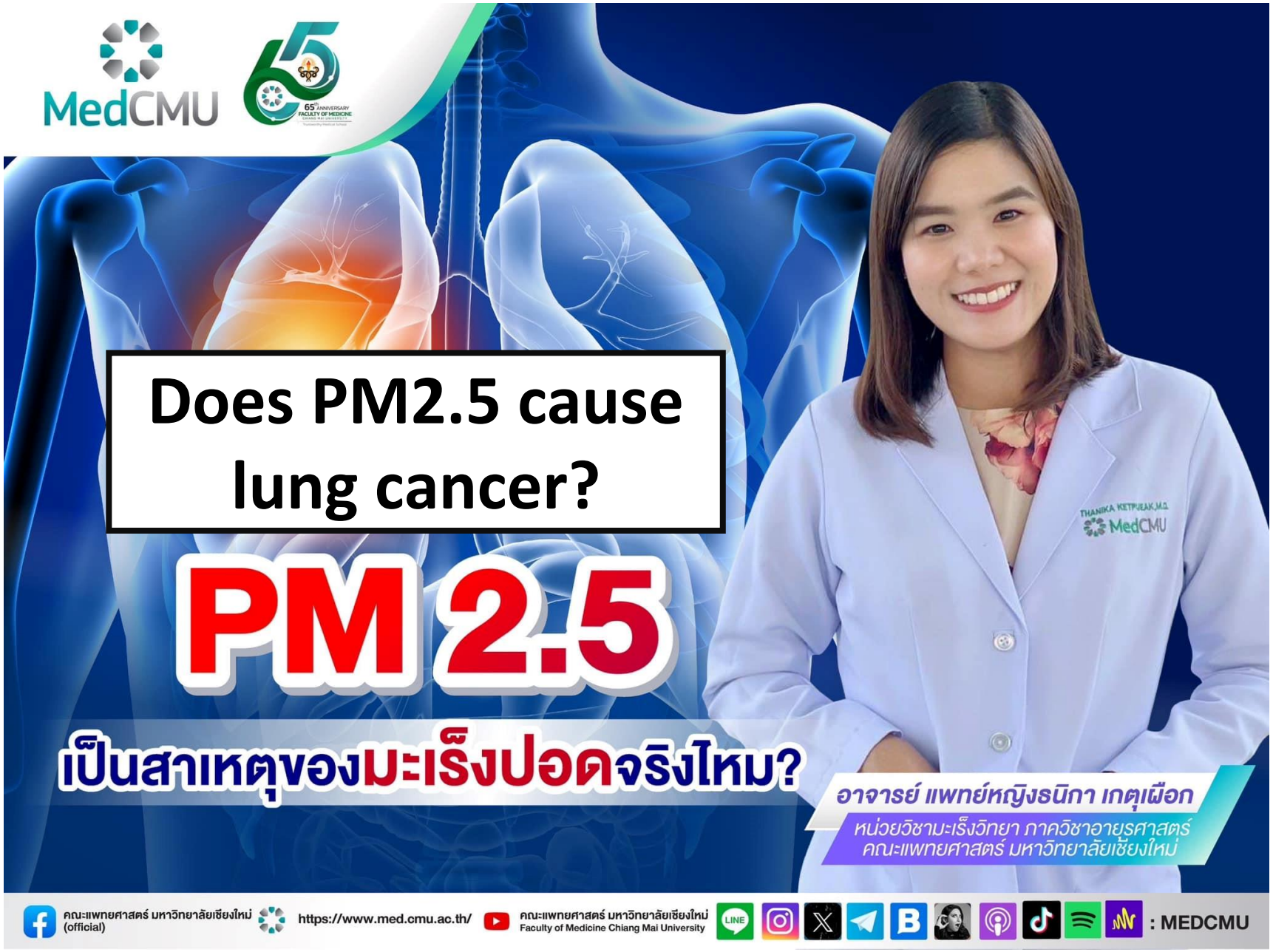
The findings in this paper do not support the conclusions that were drawn. Correlation does not imply causation.

2 19

Analisa Packham @analisapackham · Mar 6

This paper uses causal inference, my dude. Now excuse me while I go scream into a pillow.

1 70



https://www.med.cmu.ac.th/web/news-event/12683/

ปัญหามลภาวะทางอากาศของเชียงใหม่ รวมถึงหลาย ๆ จังหวัดในภาคเหนือของประเทศไทย มีมานานกว่า 10 ปี และมีความรุนแรงมากขึ้นเรื่อย ๆ โดยเฉพาะในช่วง 3-5 ปีที่ผ่านมา มลภาวะทางอากาศในระดับรุนแรงที่ดูกันง่าย ๆ จาก Air quality index (AQI) ว่าเป็นสีแดง สีม่วงนั้น ส่งผลกระทบต่อสุขภาพชัดเจน ทั้งโรคทางเดินหายใจ โรคทางหลอดเลือดสมอง และหัวใจ และแน่นอนคือมะเร็งปอด

สำหรับ Particulate matter (PM) 2.5 คือ วัตถุที่มีขนาดเล็กกว่า 2.5 ไมครอน ก็คือเล็กกว่ามิลลิเมตรพันเท่า เลยสามารถลงไปในปอดส่วนลึกได้

มะเร็งปอดคนทั่วไปจะเข้าใจว่าสัมพันธ์กับการสูบบุหรี่ เป็นความเข้าใจที่ถูกต้อง มีความสัมพันธ์กัน ยิ่งสูบเยอะ สูบนาน ยิ่งเพิ่มโอกาสในการเป็นมะเร็งปอด แต่ปัจจุบันพบว่าคนที่ไม่สูบบุหรี่สามารถเป็นมะเร็งปอดได้เช่นเดียวกันและพบได้ไม่น้อย โดยเฉพาะในคนที่มีอายุน้อย เพศหญิงและเป็นชาวเอเชียตะวันออก

เริ่มแรกตั้งแต่ปี 2009 มีการศึกษาพบว่า กลุ่มผู้ป่วยที่มีลักษณะเช่นนี้ สัมพันธ์กับยีนกลายพันธุ์ที่เรียกว่า EGFR แต่สาเหตุของการเกิดมะเร็งปอดในผู้ป่วยกลุ่มนี้นั้น มีหลายปัจจัยส่งเสริม เช่น พันธุกรรม เชื้อชาติ การได้รับสารก่อมะเร็ง และมลภาวะทางอากาศ

มีการศึกษาพบว่า ทุกๆ PM 2.5 ที่เพิ่มขึ้น 1 ไมโครกรัมมิลลิเมตร ทำให้อุบัติการณ์ของ มะเร็งปอดที่มียีนกลายพันธุ์ EGFR เพิ่มขึ้น ทั้งจากประชากรในประเทศอังกฤษ เกาหลีใต้และไต้หวัน แ

PM2.5 is a “precipitating factor” and not a cause of lung cancer. We currently cannot say that in a single patient, what exactly cause lung cancer as the process is complex and multifactorial.

ล้อมที่มีมลภ

จากกา

ปกติที่ยังไ

นอกจา

าย

พันธุ์ EGFR อยู่แล้ว และ 53% มียีนกลายพันธุ์ KRAS

ดังนั้นจากการศึกษาทั้งหมดที่กล่าวมา สรุปได้ว่า PM 2.5 เป็นปัจจัยกระตุ้นให้เกิดมะเร็งปอด โดยเฉพาะในคนที่มีการกลายพันธุ์ของยีน EGFR และ KRAS อยู่แล้ว อย่างไรก็ตามจะเห็นได้ว่า PM 2.5 นั้นเป็น “ปัจจัยกระตุ้น” ไม่ใช่สาเหตุ และในปัจจุบันเราไม่สามารถบอกได้ชัดเจนว่า ในผู้ป่วยคนหนึ่งที่เป็นมะเร็งปอดนั้นมีสาเหตุจากอะไรได้แน่นอน เนื่องจากกระบวนการเกิดมะเร็งดังกล่าวซับซ้อนและเกิดได้จากหลายปัจจัยกระตุ้นดังที่กล่าวไว้ข้างต้น

เมื่อทราบดังนี้แล้ว จึงไม่ควรอยู่ในบริเวณที่มีมลภาวะทางอากาศสูง หากหลีกเลี่ยงไม่ได้ควรใส่หน้ากากที่ป้องกัน PM 2.5 ได้แก่ หน้ากาก N95 เป็นต้นไป จึงจะสามารถกรองอนุภาคเหล่านี้ได้ งดกิจกรรมกลางแจ้ง อยู่ในอาคารและเปิดเครื่องฟอกอากาศ

Ref: https://www.nature.com/articles/s41586-023-05874-3

บทความโดย : อาจารย์ แพทย์หญิงธนิกา เกตุเผือก หน่วยวิชามะเร็งวิทยา ภาควิชาอายุรศาสตร์ คณะแพทยศาสตร์ มหาวิทยาลัยเชียงใหม่

Big Data, Data Science, and Causal Inference: A Primer for Clinicians

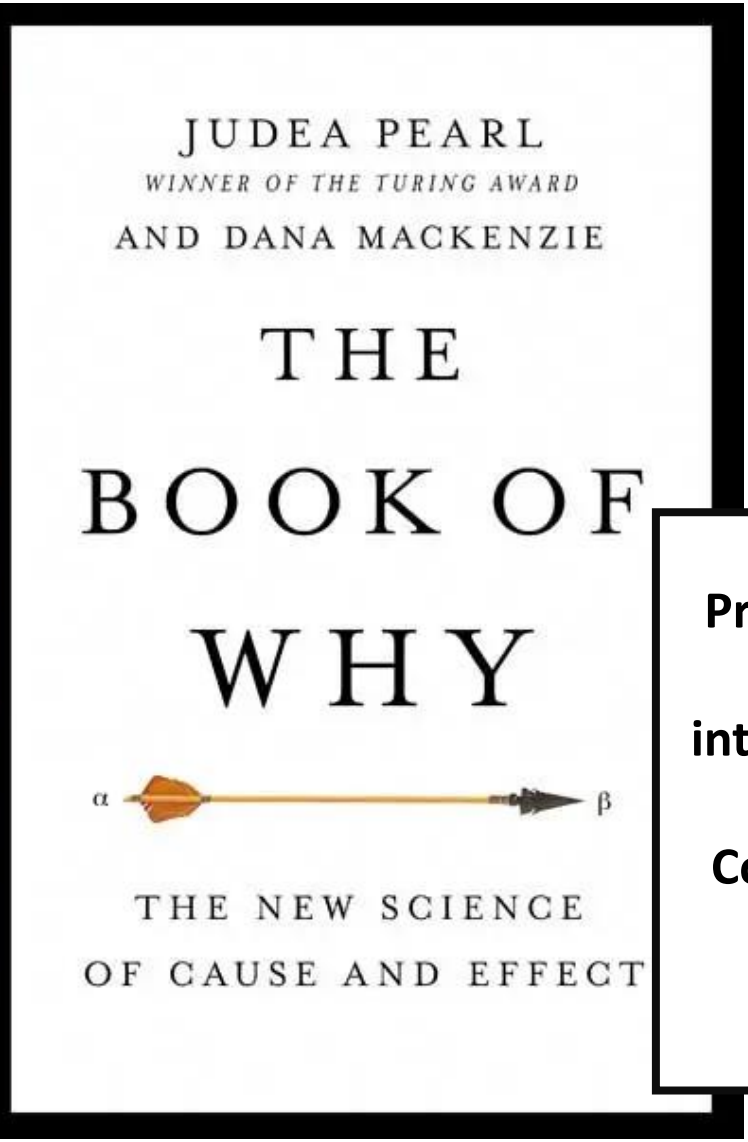
Yoshihiko Raita [1], Carlos A. Camargo Jr. [1,2,3], Liming Liang [1,3,4] and Kohei Hasegawa [1,3,4]*

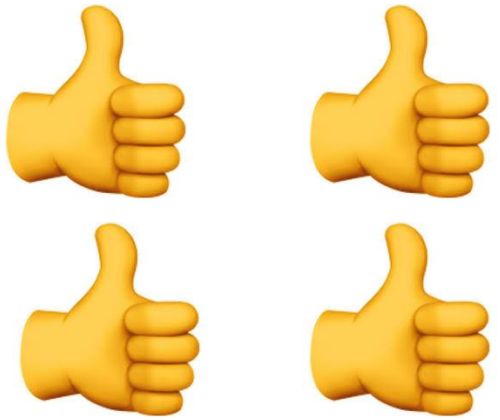
[1] Department of Emergency Medicine, Harvard Medical School, Massachusetts General Hospital, Boston, MA, United States, [2] Division of Rheumatology, Allergy, and Immunology, Department of Medicine, Harvard Medical School, Massachusetts General Hospital, Boston, MA, United States, [3] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, United States, [4] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States

Nat Sirirutbunkajorn
Radiation oncologist, Ramathibodi hospital
Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol university
Nut19012537@gmail.com

Goal of data Science and the Ladder of Causation

- It is important to understand **what data science is** (and is not).
- Thus, we organize questions and task of data science according to the **Ladder of Causation**.





8.5/10
Pro: Easy to understand. Not too math heavy, good introduction to causality from the past to present.
Con: Historical part can be a slog (but important for understanding)
- Nat Sirirutbunkajorn

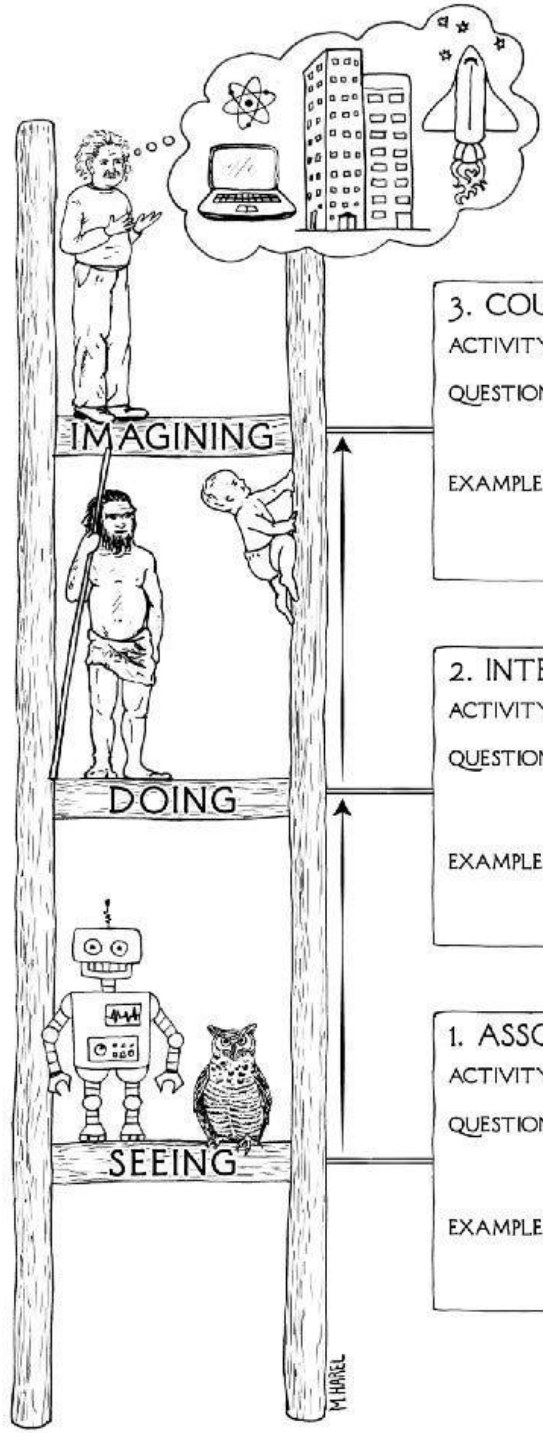
TABLE 2 | Scientific questions, required information, and analytical methods of data science according to the ladder of causation.

	Examples of scientific question	Required information*	Examples of analytics and tools
Rung 1 association and prediction	- What are the risk factors for developing asthma? - What is the probability of developing asthma in a patient with a set of predictors?	- Risk factors/predictors - Outcomes	- Regression - Supervised machine learning algorithms (e.g., random forests, neural network/deep learning)
Rung 2 intervention	Will a new biologic agent decrease the rate of asthma exacerbation by 30%, compared to placebo?	- Eligibility criteria - Exposures/treatments - Outcomes	- Elementary statistics in RCTs (e.g., risk differences of the outcome) - Intention-to-treat analysis - Per-protocol analysis - Causal Bayesian network
Rung 3 counterfactual causal inference	What would be the preventive effect of a new drug had it been given to a group of patients with a set of characteristics?	- Eligibility criteria - Exposures/treatments - Outcomes - Observation period and temporality† - Domain knowledge on the causal structure (e.g., confounders, mediators, colliders)	- Regression - Propensity score matching - Standardization/G-formula - IPW/MSM - Targeted learning - IV-methods/Mendelian randomization

IPW/MSM, inverse probability weighing for marginal structure model; IV, instrumental variable; RCT, randomized controlled trial.

**For all tasks, no information bias (no measurement error or misclassification) and no model misspecification are required.*

†*The effect of interest must occur after the cause (and an expected delay) during an observation period.*



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

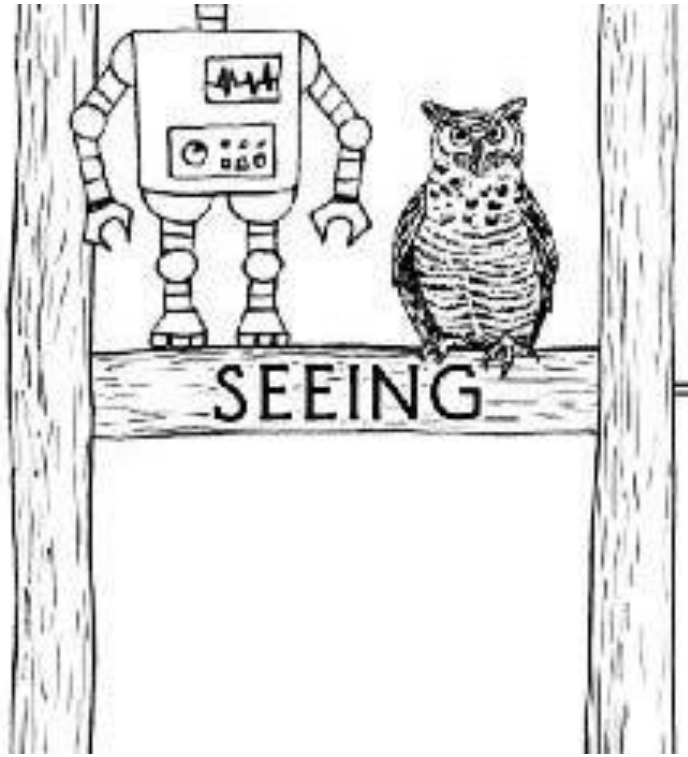
1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symprom tell me about a disease? What does a survey tell us about the election results?

Rung 1: Association and Prediction



1. ASSOCIATION

ACTIVITY:	Seeing, Observing
QUESTIONS:	*What if I see ...?* (How are the variables related? How would seeing X change my belief in Y?)
EXAMPLES:	What does a symptom tell me about a disease? What does a survey tell us about the election results?

Rung 1: Association and Prediction

- **Association** invokes **exclusively probabilistic relationships** between the variables within observed data.
  - "Recurrent wheezing in early childhood is associated with the development of asthma"
    - The **probability of observing one variable depends on that of the other** (or vice versa)
- **Prediction** maps the derived probabilistic association to **future data** in order to **forecast the conditional probability of outcome**.
  - e.g. clinical risk score (Ex. Asthma predictive index) or polygenic risk score.
- Tools used:
  - Basic computation/traditional statistics (e.g. regression models)
  - **ML, deep learning**
    - excels in association and prediction task but **lack causal reasoning**

Rung 1: Association and Prediction

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

Paul Koch
Microsoft Research
paulkoch@microsoft.com

Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

https://dl.acm.org/doi/10.1145/2783258.2788613

Rung 1: Association and Prediction

- **In the mid 90's,** a large multi-institutional project was funded by Cost-Effective HealthCare (CEHC) **to evaluate ML in healthcare such as predicting pneumonia risk.**
- Goal was to **predict the probability of death (POD) patients with pneumonia**
  - high-risk patients **could be admitted to the hospital**
  - low-risk patients **were treated as outpatients.**
- TLDR; neural nets won (AUC=0.86) but they were considered too risky and instead logistic regression was chosen.

https://www.youtube.com/watch?v=wql_z1yumzY&list=PLUl4u3cNGP60B0PQXVQyGNdCyCTDU1Q5j&index=6

Rung 1: Association and Prediction

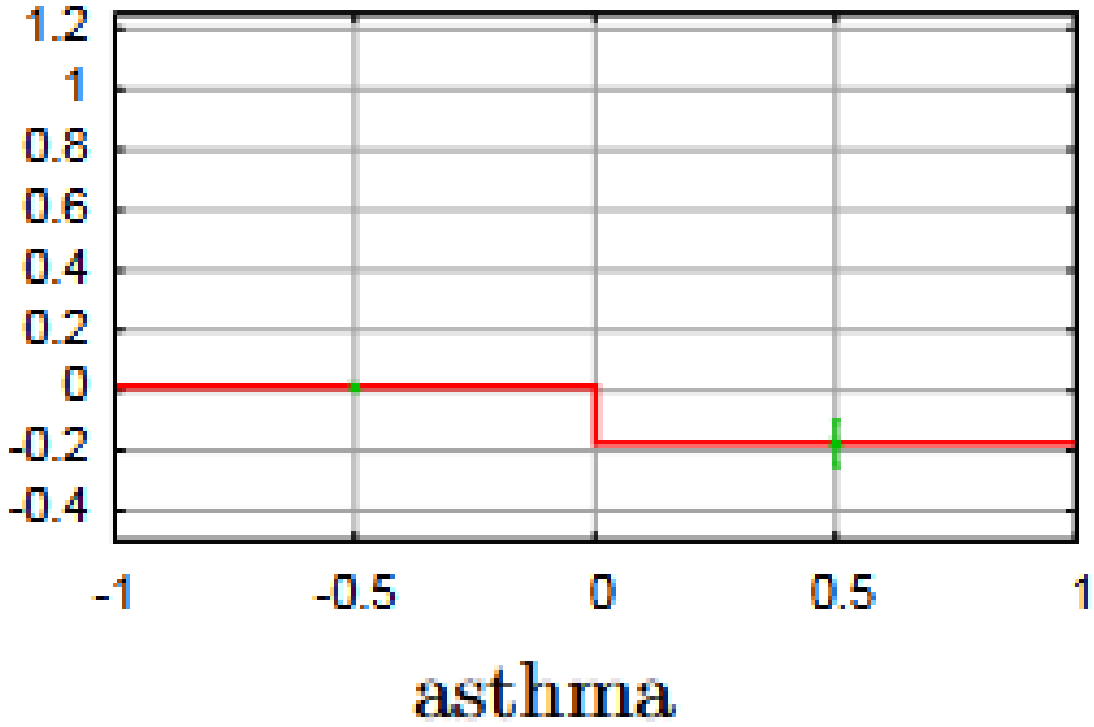
Label = Death

Model	Pneumonia	Readmission
Logistic Regression	0.8432	0.7523
GAM	0.8542	0.7795
GA^2M	0.8576	0.7833
Random Forests	0.8460	0.7671
LogitBoost	0.8493	0.7835

Table 2: AUC for different learning methods on the pneumonia and 30-day readmission tasks.

https://dl.acm.org/doi/10.1145/2783258.2788613

Rung 1: Association and Prediction



Having Asthma reduce the predicted probability of death!

https://dl.acm.org/doi/10.1145/2783258.2788613

Rung 1: Association and Prediction

- Rule-based system learned the rule **"HasAsthama(x) ⇒ LowerRisk(x)"**
  - **Patient with asthma -> admit directly to ICU -> receive aggressive care -> lower risk of overall death!**
- If the rule-based system had learned that asthma lowers risk, certainly, the neural nets had learned it, too.
  - **Neural net was not used because lack of intelligibility made it difficult to know what other problems might also need fixing**

- Formally, this is what's happening:

A long survival time may be because of treatment!

Rung 1: Association and Prediction

5.7 Correlation Does Not Imply Causation

Because the models in this paper are intelligible, it is tempting to interpret them causally. Although the models accurately explain the predictions they make, they are still based on correlation. If features were added to or subtracted and the model retrained, the graphs for some terms that had remained in the model would change because of correlation with the features added or subtracted. Although details of some of the shape plots are suggestive (e.g., does pneumonia risk truly jump as age increases above 65, and again above 85?), it is not (yet) clear if some details like this are due to a) overfitting; b) correlation with other variables; c) interaction with other variables; d) correlation or interaction with unmeasured variables; or e) due to true underlying phenomena such as retirement and change in insurance provider.

Perhaps the strongest statement we can make right now is that the models are intelligible enough to provide a window into the data and prediction problem that is missing with many other learning methods, and that this window allows questions to be raised that will require investigation and further data analysis to answer. In future versions of these models we hope to automate some of these analyses so that it is clearer what features in the intelligible model are "real" or due to random factors such as overfitting and spurious

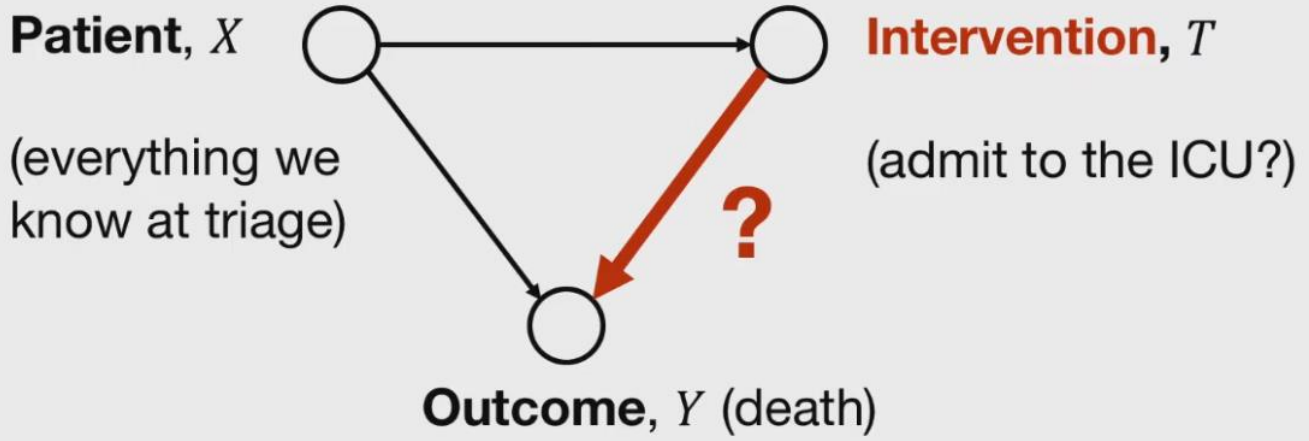
Intelligible model (or explainable AI) makes it easy to tell what drive the prediction, but it should not be interpreted casually

https://dl.acm.org/doi/10.1145/2783258.2788613

Rung 1: Association and Prediction

Intervention-tainted outcomes

- The rigorous way to address this problem is through the language of **causality:**



https://dl.acm.org/doi/10.1145/2783258.2788613

Rung 1: Association and Prediction

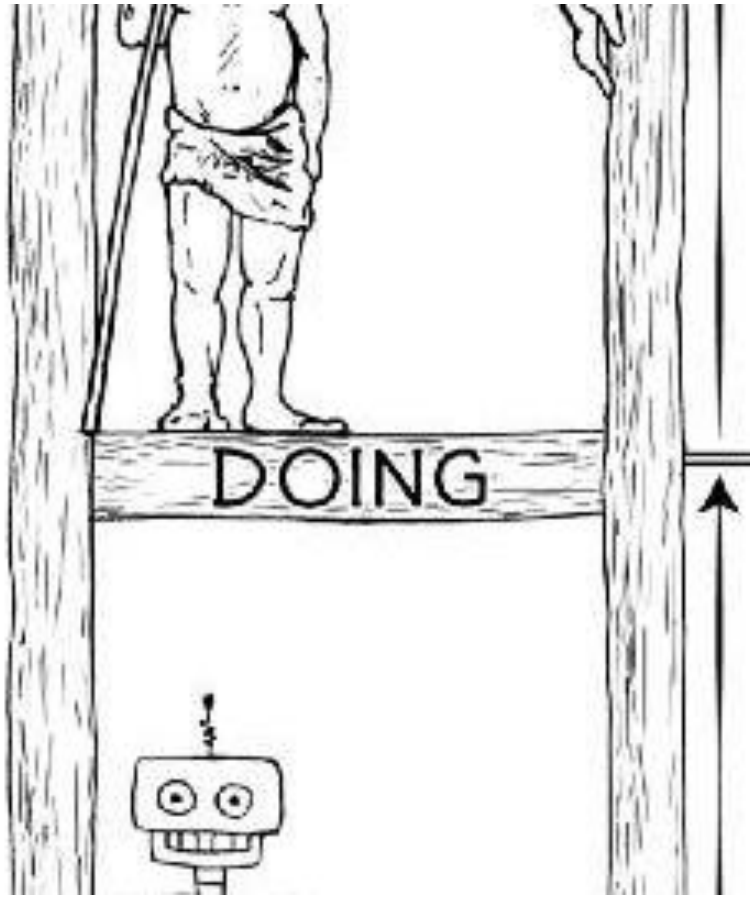
- **In the mid 90's,** a large multi-institutional project was funded by Cost-Effective HealthCare (CEHC) **to evaluate ML in healthcare such as predicting pneumonia risk.**
- Goal was to **predict the probability of death (POD) patients with pneumonia** so that
  - high-risk patients **could be admitted to the hospital**
  - low-risk patients **were treated as outpatients.**
- TLDR; neural nets won (AUC=0.86) but they were considered too risky and instead logistic regression was chosen.

What rung is this objective/task?

Do the tools they used make sense for the task?

https://dl.acm.org/doi/10.1145/2783258.2788613

Rung 2: Intervention



2. INTERVENTION

ACTIVITY: Doing, Intervening

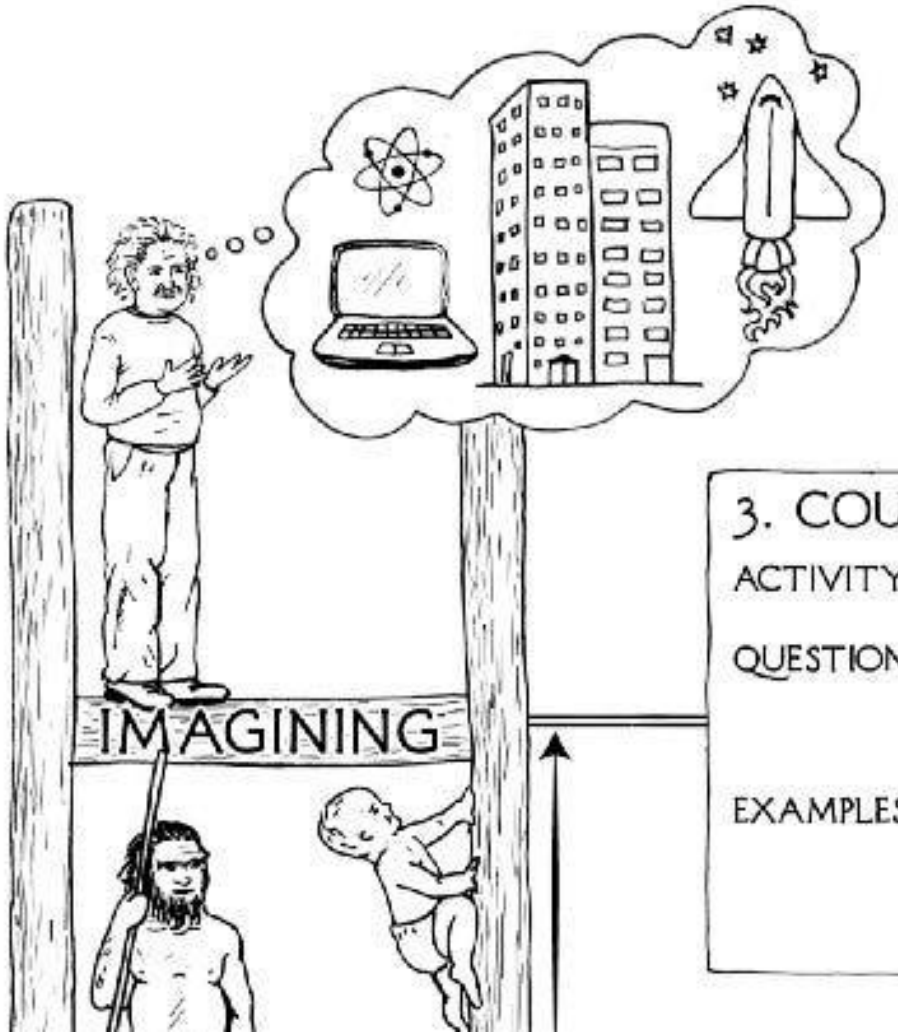
QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

Rung 2: Intervention

- **Intervention** involves not only observing the data but also **changing what we observe according to our causal hypothesis.**
- **RCTs** which meets some assumption has been considered the **goal standard.**
  - **e.g. The average causal effect of drug X and mortality Y is 0.5.**
- However, **no experiment cannot handle a "what if?" question.**
  - "what if this patient had received treatment X at time t?"

Rung 3: Counterfactual



3. COUNTERFACTUALS

ACTIVITY:	Imagining, Retrospection, Understanding
QUESTIONS:	*What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES:	Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

Rung 3: Counterfactual

- **What would happen had y happened/not happened?**
  - **e.g. For patient who receive treatment and died how many would not die had they not receive treatment: $P(Y0=0|X=1,Y=1)$**
- Rung 3 requires more information than rung 2 to answer.
  - e.g. Consider an RCT which the drug's average causal effect is 0.

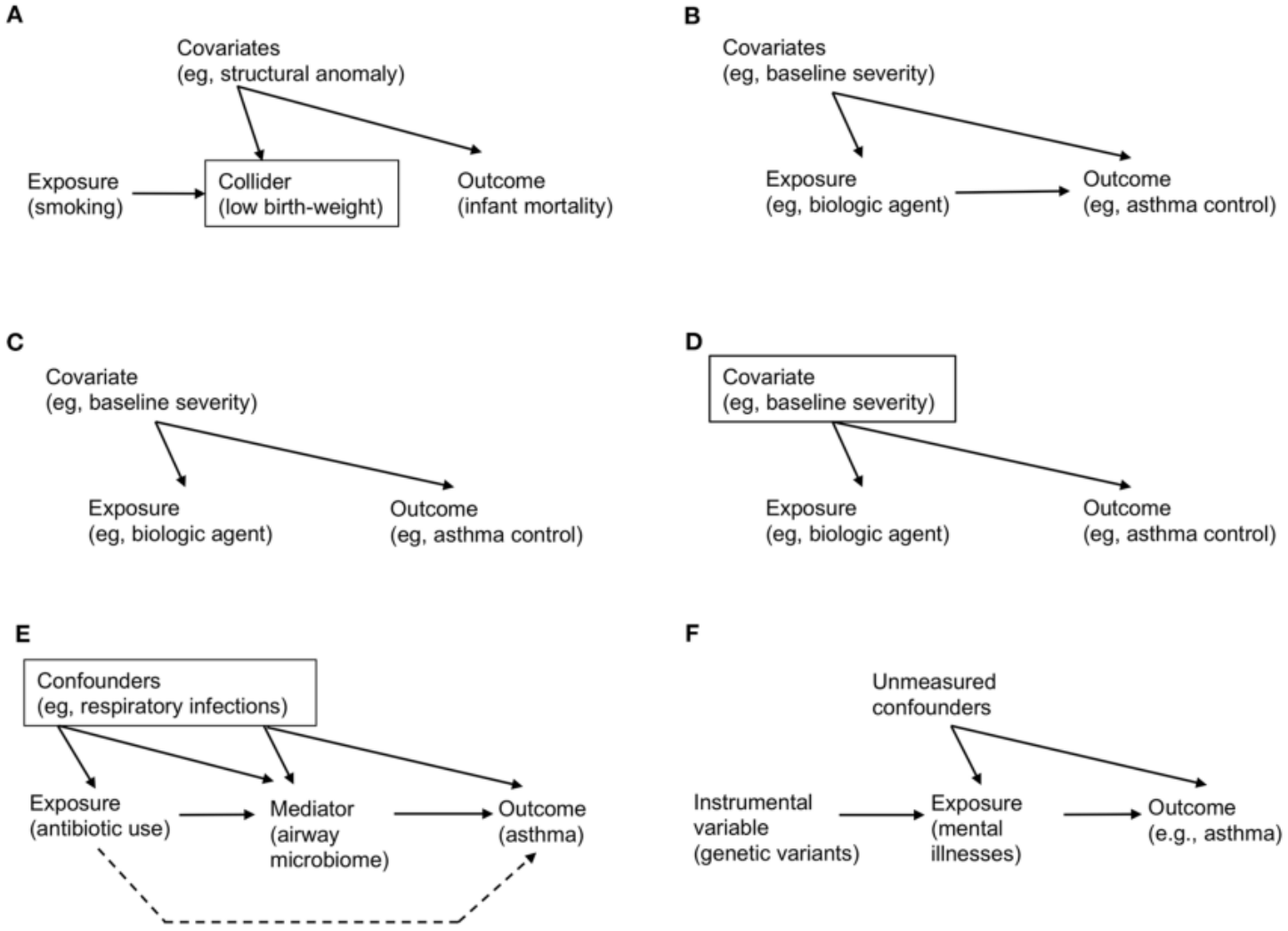
Model 1	$u_2 = 0$		$u_2 = 1$		Marginal	
	$x = 1$	$x = 0$	$x = 1$	$x = 0$	$x = 1$	$x = 0$
$y = 1$ (death)	0	0	0.25	0.25	0.25	0.25
$y = 0$ (recovery)	0.25	0.25	0	0	0.25	0.25

Model 2	$u_2 = 0$		$u_2 = 1$		Marginal	
	$x = 1$	$x = 0$	$x = 1$	$x = 0$	$x = 1$	$x = 0$
$y = 1$ (death)	0	0.25	0.25	0	0.25	0.25
$y = 0$ (recovery)	0.25	0	0	0.25	0.25	0.25

u = some factor which cause treatment heterogeneity

https://stats.stackexchange.com/questions/379799/difference-between-rungs-two-and-three-in-the-ladder-of-causation

Major Causal Inference Tools



The way forward

- In medicine, is important to remember that a **data-driven algorithm** may excel at predicting but **is agnostic about the reason and possible measures to have prevented it.**
- **Identifying patients with a worse prognosis** (through prediction) is a different question from **identifying the optimal prevention and treatment strategies** for a specific group of patients—the defining question of **precision medicine** (through causal inference).
- Casual structure usually is unknown and most researches tended to answer relatively narrow causal question (e.g., the average treatment effect of bronchodilators in infants with bronchiolitis).
- Integration of bigdata with data science approaches could help in these tasks for optimal treatment decision making.

TABLE 4 | Twelve major resources for clinicians who wish to learn about data science.

Topic	Type	Platform/Resource	Content summary
Data science (in general)	MOOC	Kahn academy	An online course that covers a wide range of topics about statistical analyses
	MOOC	Coursera: data science specialization	An online course that provides a broad overview of data science
	MOOC	edX: introduction to probability (HarvardX STAT110x)	An online course that introduces the basics of probability theories, which are fundamental for data science, statistics, and causal inference
	MOOC	Stanford: statistical learning	An online learning course that offers an introduction to various statistical learning (including machine learning) approaches
	Textbook	*An Introduction to Statistical Learning*	A well-written introductory textbook that is used in the statistical learning course (see above)
	Paper	*BMJ*: research methods & reporting	*BMJ* series introduces important topics of epidemiology and biostatistics to help clinicians interpret the medical literature
	Paper	*JAMA*: guide to statistics and medicine	*JAMA* series introduces important statistical techniques to help clinicians interpret the medical literature
Machine learning	MOOC	Coursera: machine learning	One of the most popular machine learning courses (as of January 2021, 3.9 million students have been enrolled). This introductory course provides an overview of various machine learning algorithms
	MOOC	Coursera: Deep learning specialization	A more detailed online course that covers the basics and applications of various deep learning algorithms
Causal inference	MOOC	edX: Causal diagrams (HarvardX PH559x)	An online course that introduces an overview of causal diagrams in clinical research
	MOOC	Coursera: A crash course in causality	An online course offered that provides an introductory overview of causal inference theories and approaches
	Textbook	*Causal Inference in Statistics: A Primer* (64)	Introductory-level textbook that covers important topics in causal inference (e.g., causal diagram)
	Textbook	*Causal Inference: What if* (15)	Comprehensive intermediate-level textbook that provides the concepts of and methods for causal inference in clinical research
Programming	MOOC	Coursera: foundations using R specialization	An online course that provides a broad overview of R programing
	Others	DataCamp	A collection of introductory video lectures and hand-on coding practices in several programing languages (e.g., R, python)

MOOC, massive open online course; BMJ, British Medical Journal; JAMA, Journal of the American Medical Association.
All of the listed MOOCs are publicly-available without fee.