The Cycle of Statistical Research

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

February 19, 2020 @ CCIMI Seminar, Cambridge

Slides and more information are available at
`http://www.statslab.cam.ac.uk/~qz280/`.

About me

- "New" University Lecturer in the Stats Lab.
- PhD (2011-2016) in Statistics from Stanford.
- Postdoc (2016-2019) at University of Pennsylvania.
- Current research area: **Causal Inference**.
- Interested applications: public health, genetics, social sciences, computer science.

Growing interest in causal inference

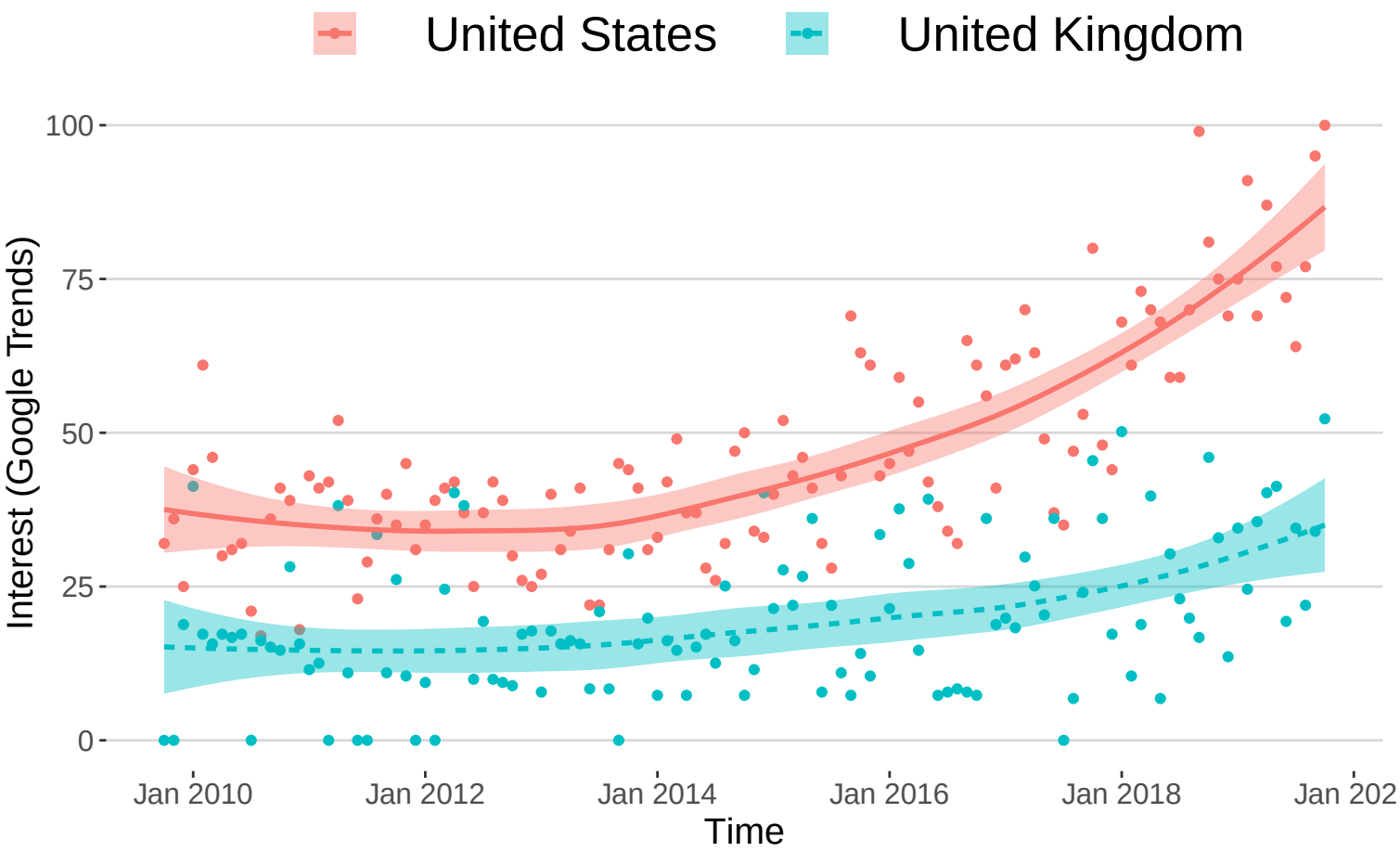


Figure: Data from Google Trends.

Why study causal inference?

Old and new problems

- Epidemiology and public health: effectiveness of prevention/treatment, causal effect of risk factors, etc.
- Quantitative social sciences: evaluation of social programs, policy impact, etc.
- Precision medicine.
- Massive online experiments.
- Explanation and fairness of machine learning algorithms.

From casual inference to causal inference

Understanding causal inference provides us a comprehensive **cyclic view of statistical research**.

Statistics vs. Data Analysis

Buzzwords

- Data mining;
- Machine learning;
- Big data;
- Data science;
- Artificial intelligence;
- Mathematics of information ☺

A much older love-hate relationship

Statistics and **Data Analysis**

Statistics

Definitions

- Broader: **"the science of using information discovered from studying numbers"** (Cambridge Dictionary).
- Narrower: **"the application of probability theory, a branch of mathematics, to statistics, as opposed to techniques for collecting statistical data"** (Wikipedia for mathematical statistics).

History

Three movements:

- Around 1900: Standard deviation, correlation, regression analysis, method of moments, χ^2-test, student's t-test, ... (Galton, Pearson, Gosset, ...).
- 1920s – 1930s: Hypothesis testing, sufficient and ancillary statistics, Fisher information, randomised experiments and experimental design (Fisher).
- 1930s – 1940s: Confidence intervals, power of a statistical test, stratified sampling, decision theory (Pearson, Neyman, Wald, ...).

Data Analysis

The future of data analysis (Tukey, 1961a)

For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, ... it has become clear that **their "dealing with fluctuations" aspects are ... of lesser importance than ... to deal effectively with the simpler case of very extensive data**, *where fluctuations would no longer be a problem.*

I have come to feel that my central interest is in data analysis, ... : **procedures** *for analysing data,* **techniques for interpreting** *the results of such procedures,* **ways of planning the gathering of data** *to make its analysis easier, more precise or more accurate, and all the* **machinery and results of (mathematical) statistics** *which apply the analysing data.*

Tukey is known for

- Coining the term "bit";
- Co-inventing the Fast Fourier Transform (FFT) algorithm;
- Tukey range test and later developments on Multiple comparisons;
- Developing a variety of data visualisation tools (boxplot, projection pursuit, Tukey median and Tukey depth);
- Advocating for "exploratory data analysis".

Danger with data analysis (and data science)

Presidential Address to the American Statistical Association (Box, 1979)

Please can Data Analysts get themselves together again and become whole Statisticians before it is too late? *Before they, their employers, and their clients forget the other equally important parts of the job statisticians should be doing, such as designing investigations and building models?*

By invention of the concept of **Experimental Design**, *Fisher promoted the statistician from a curator of dusty relics to a valued member of a scientific team,* **responsible for planning and taking part in the conduct of an investigation**. *Let us not allow him to be relegated to his previous passive and inferior role by an injudicious choice of a name, "Our Data Analyst" is too close for my liking to "Our Tame Statistician", a poor thing if that is all he is.*

Box is known for

- "All models are wrong, but some are useful";
- Box-Cox transformation;
- His work on experimental design.

(Box married a daughter of Fisher's.)

Statistics vs. Data Analysis: A love-hate relationship

My translation

- Tukey: Statistical research is not just about proving mathematical theorems, but also about how to deal with real data.
- Box: Statistical research is not just about doing what we are told by our supervisors or clients, but also about bringing thoughts and rigour to scientific investigations.

Tukey and Box actually shared (almost) the same sentiment!

The only difference is that they were attacking different narrow-minded views:

- Tukey was worried about the **mathematical view** of statistical research becoming dominant, so he emphasised on the **algorithmic view**.
- Box was worried about the **algorithmic view** of statistical research becoming dominant, so he emphasised on the **mathematical view**.

The cycle of statistical research

Tukey (1961b) quoting Box (1957)

But if an oversimple paradigm is to be selected, George Box's recent expression of the situation will serve excellently. He says: " Scientific research is usually an iterative process. **The cycle: conjecture–design–experiment–analysis leads to a new cycle of conjecture–design–experiment–analysis and so on**.... *The experimental environment ... and techniques appropriate for design and analysis tend to change as the investigation proceeds."*

Tukey (1961b)'s question

The research problem involving statistical and quantitative methodology . . . is a problem in higher education and in the cultural anthropology of scientists: **Why do so few learn to analyse data well?**

Tukey suggested that the solution is to **let Ph.D. students to go through all the phases of the cycle**. Has this been implemented after nearly 60 years?

Rest of the talk

1. How causal inference can help us to gain a cyclic view of statistical research.
2. Example 1: the Lipid Hypothesis.
3. Example 2: the epidemic growth of the COVID-2019 outbreak.

Causality

Goals of statistical research

- **Description of a population**: **1%**;
- **Predicting the response of another sample**: **9%**;
- **Understanding the causal relationship between variables**: **90%**
  (although most wouldn't say the word "causal", for reasons in the next slide).

Randomisation

The breakthrough

- In 1920s, Fisher first introduced randomisation as a principled way to establish causality in scientific research (*The Design of Experiments*, 1935).
- The idea dates back to the philosopher Perice in the late 1800s.

The narrow-minded view of causality

- **"Correlation does not imply causation"**
- $\Longrightarrow$ **Causality can only be established by randomised experiments**
- $\Longrightarrow$ **Experimental design = Improve the efficiency.**
- Example: "Use of Causal Language" in the author guidelines of *JAMA*:

> **Causal language** *(including use of terms such as effect and efficacy)* **should be used only for randomised clinical trials**. *For all other study designs (including meta-analyses of randomised clinical trials), methods and results should be described in terms of association or correlation and should avoid cause-and-effect wording.*

- Statistical research is **a chain**: conjecture $\rightarrow$ design $\rightarrow$ experiment $\rightarrow$ analysis

"Clouds" over randomised experiments

(Borrowing the metaphor from the famous 1900 speech by Kelvin.)

Smoking and Lung cancer (1950s)

- Hill, Doll and others: **Overwhelming association** between smoking and lung cancer, **in many populations**, and **after conditioning on many variables**.
- Fisher and other statisticians: **But correlation is not causation.**

Infeasibility of randomised experiments

- Ethical problems, high cost, and other reasons.

Non-compliance

- People may not comply with assigned treatment or drop out during the study.

How to define causality?

Definition 0: Implicitly from randomisation

If people were randomised to take one of two treatments (binary variable A), and all other characteristic are (stochastically) the same, then any difference in the outcome Y *must* be caused by the different treatments.

Definition 1: Potential outcome (Neyman, 1923; Rubin, 1974)

People have two potential outcomes (also called counterfactuals), $Y(0)$ and $Y(1)$. We only observe one counterfactual, $Y = Y(A) = AY(0) + (1 - A)Y(1)$, but we would like to infer about the difference between

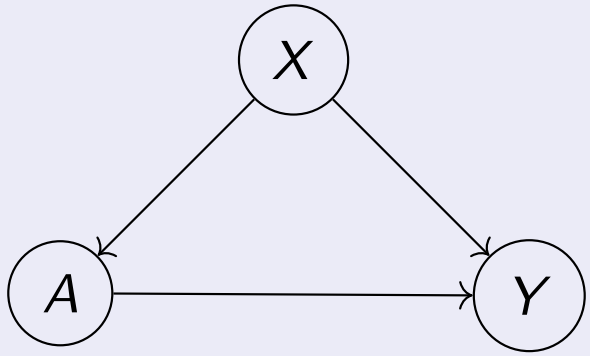
$$\text{Distribution of } Y(0) \text{ vs. Distribution of } Y(1).$$

- How is this possible? If we know $A \perp\!\!\!\perp Y(0) \mid X$, then

$$\begin{aligned}\mathbb{P}(Y(0) = y) &= \mathbb{E}[\mathbb{P}(Y(0) = y \mid X)] \\ &= \mathbb{E}[\mathbb{P}(Y(0) = y \mid A = 0, X)] \\ &= \mathbb{E}[\mathbb{P}(Y = y \mid A = 0, X)]\end{aligned}$$

How to define causality?

Definition 2: Graphical model



- Bayesian networks/probabilistic graphical models (Pearl, 1985; Lauritzen, 1996): Joint distribution factorises according to the graph:

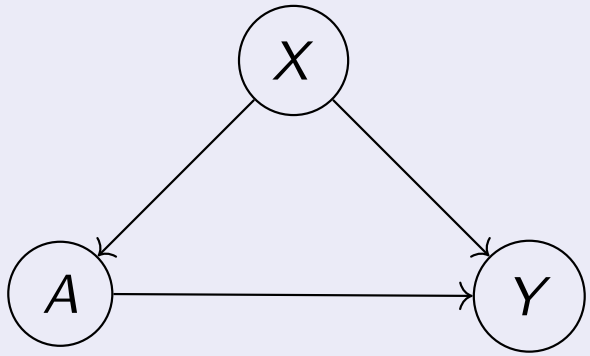
$$\begin{aligned}&\mathbb{P}(A=a, X=x, Y=y)\\ =&\mathbb{P}(X=x)\,\mathbb{P}(A=a \mid X=x)\,\mathbb{P}(Y=y \mid X=x, A=a).\end{aligned}$$

- Causal graphical models (Robins, 1986; Spirtes et al., 1993; Pearl, 2000): Joint distribution in interventional settings also described by the graph:

$$\begin{aligned}&\mathbb{P}(X=x, A=a, Y(a)=y)\\ =&\mathbb{P}(X=x)\,\mathbb{P}(A=a \mid X=x)\,\mathbb{P}(Y(a)=y \mid X=x).\end{aligned}$$

How to define causality?

Definition 3: Structural equations (Wright, 1920s; Haavelmo, 1940s)



- From the graph we may define a set of structural equations:

$$X = f_X(\epsilon_X),$$
$$A = f_A(X, \epsilon_A),$$
$$Y = f_Y(A, X, \epsilon_Y).$$

- Parameters in the structural equations are **causal effects**. For example, if $f_Y(A, X, \epsilon_Y) = \beta_{AY}A + \beta_{XY}X + \epsilon_Y$, then β_{AY} is the causal effect of A on Y.

Unification of the definitions

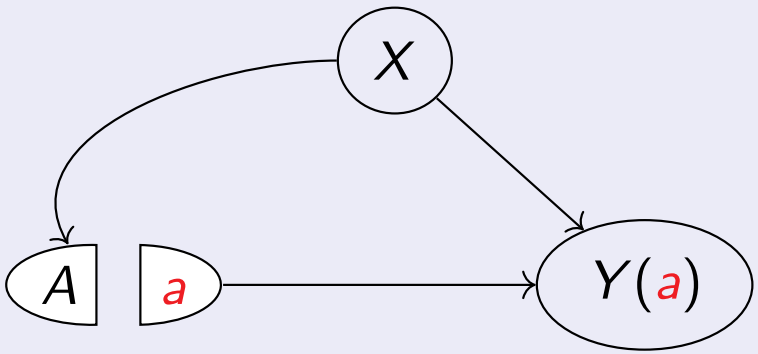
Define counterfactual from graphs

- Structural equations are **structural** instead of **regression** because they also govern the interventional settings (Pearl, 2000):

$$Y(a) = F_Y(a, X, \epsilon_Y).$$

Implied graph for counterfactuals

- Distribution of counterfactuals factorises according to an implied graph, obtained by splitting and relabelling the nodes (Richardson and Robins, 2013).



Modern causal inference

Strengths of the different approaches

- Graphical model: Good for understanding the scientific problems.
- Structural equations: Good for fitting simultaneous models for the variables.
- Counterfactuals: Good for articulating the inference for a small number of causes and effects.

The broader view of causality

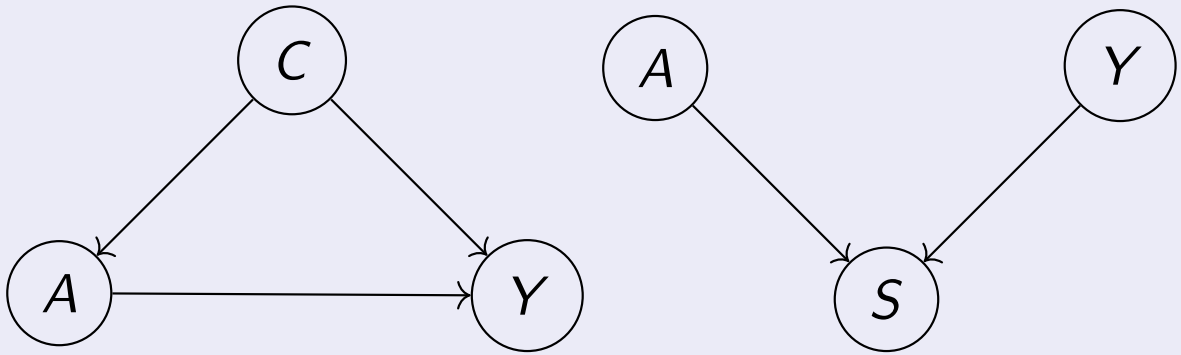
- Causality can be established from non-randomised studies, given **strong unverifiable assumptions**.
- Example: we can never test $A \perp\!\!\!\perp Y(a) \mid X$ using empirical data because we only observe $Y(a)$ for one a. (In other settings we may falsify some assumptions but can never verify it.)
- **Strength of causal inference = credibility of the assumptions.**
- Example: $A \perp\!\!\!\perp Y(a) \mid X$ is safe in a randomised experiment.
- Statistical research becomes **a cycle**: conjecture $\rightarrow$ design $\rightarrow$ experiment $\rightarrow$ analysis $\rightarrow$ conjecture $\rightarrow$....

Why is causal inference essential for this cyclic view?

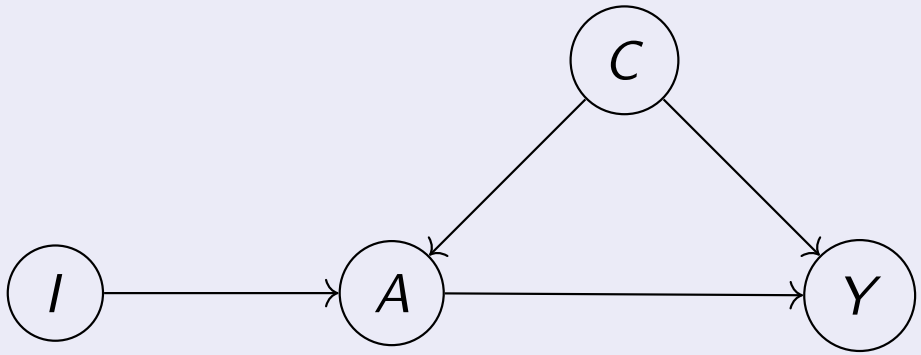
It forces us to think about the underlying **data generating mechanism** and **how to collect data**.

Key concepts can be formalised in causal inference

- Confounding (not observing C) and selection bias (conditioning on S):



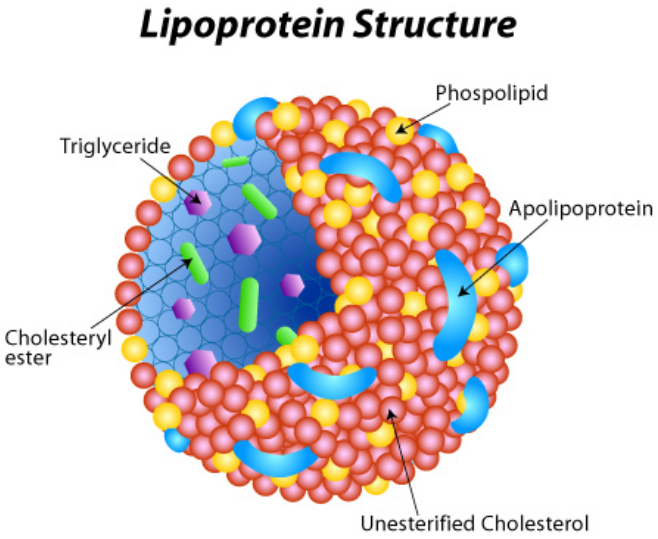
- Instrumental variable (I) and causal mechanism (no direct effect of I on Y):

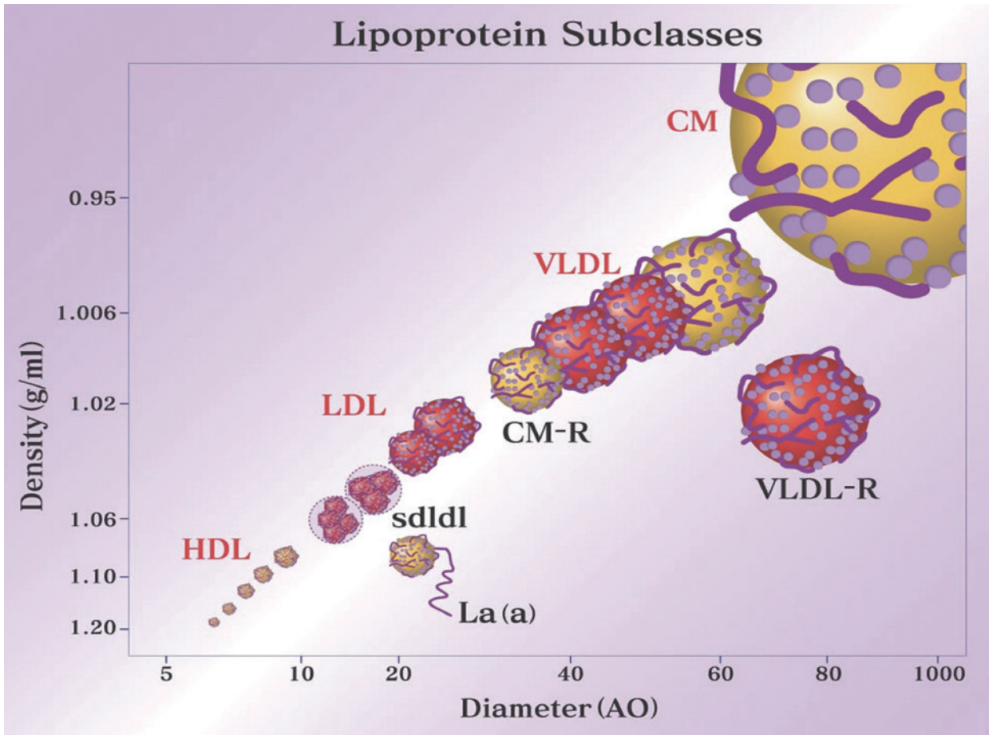


Rest of the talk

1. Example 1: the Lipid Hypothesis.
2. Example 2: the epidemic growth of the COVID-2019 outbreak.

Some background about blood lipids





Left: Lipoprotein particles transport fat molecules in our body.[1]

Right: They can be categorised based on density and size.[2]

[1] `https://www.labce.com/spg659279_lipoprotein_particles.aspx`.

[2] Nakajima, K. "Remnant Lipoproteins: A Subfraction of Plasma Triglyceride-Rich Lipoproteins Associated with Postprandial Hyperlipidemia." *Clinical & Experimental Thrombosis and Hemostasis* 1.2 (2014): 45-53.

Example 1: The Lipid Hypothesis

"Decreasing blood cholesterol significantly reduces the risk of cardiovascular diseases."

History

1913	First evidence from a **rabbit study**.
1950s – 1980s	Accumulation of evidence from **observational studies**. Transformation to **the LDL hypothesis**.
1970s	Discoveries of the biological regulation of LDL cholesterol $\rightarrow$ Brown and Goldstein winning the Nobel prize in 1985.
1980s	More evidence from **US Coronary Primary Prevention Trial**.
1990s	Scepticism continued until landmark **statin trials**.
2010s	Reaffirmation from **Mendelian randomisation**.

However, the role of **HDL cholesterol** remains quite controversial.

The HDL Hypothesis

"HDL is protective against heart diseases."

History

	1960s	Formulation of the hypothesis from **observational studies**. The inverse association has been firmly established over the years.
	1980s	Supporting evidence from **animal studies**.
But...	2000s	Null findings from **studies of Mendelian disorders**.
	2010s	**Failed randomised trials** using *CETP* inhibitors (*CETP* is an enzyme responsible for moving cholesterol from HDL particles to LDL particles).
	2010s	**Null findings from Mendelian randomisation**.

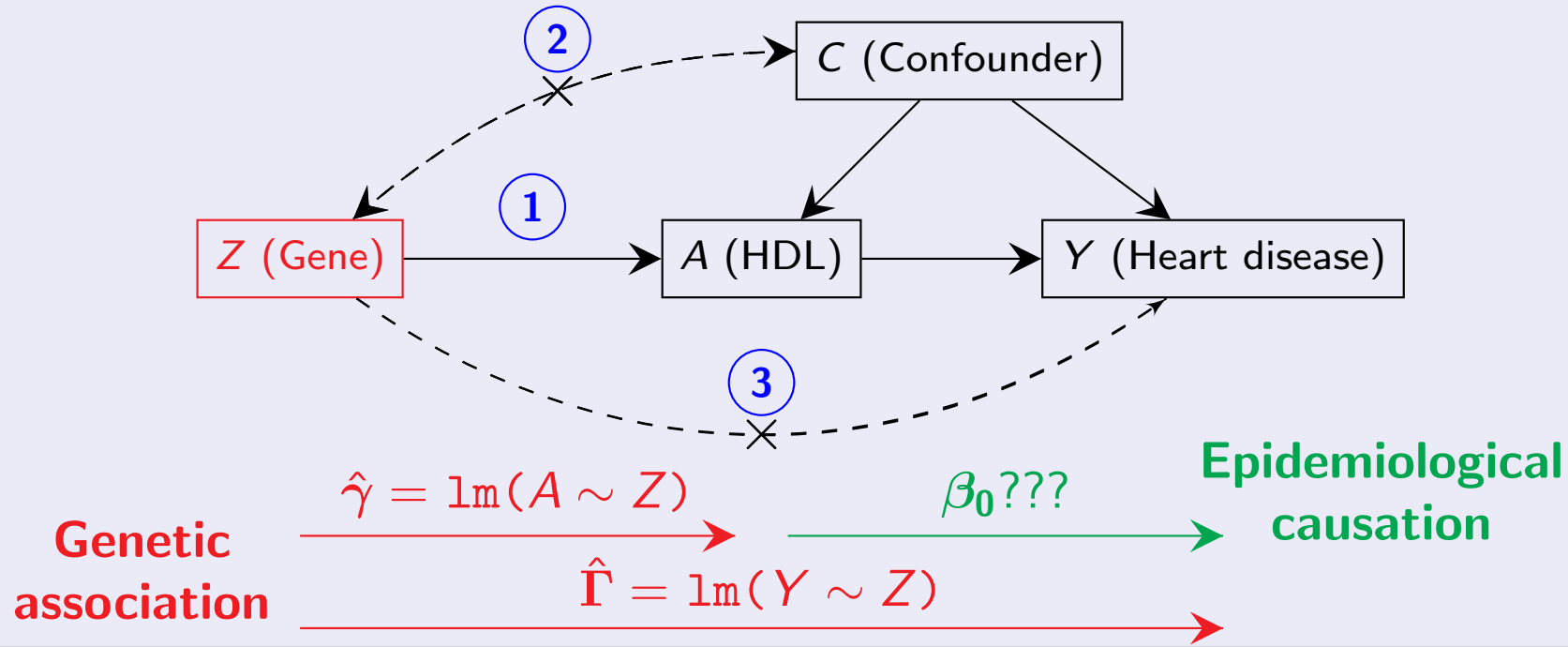
New York Times article reporting an article published in *Lancet* (May 16, 2012),:

> *"I'd say* **the HDL hypothesis is on the ropes right now**," *said Dr. James A. de Lemos . . .*
>
> *Dr. Kathiresan said. "I tell them, ' It means you are at increased risk, but* **I don't know if raising it will affect your risk.**"'

What did the *Lancet* article (Voight et al., 2012) do?

Mendelian randomisation

Using genetic variation as instrumental variables:



What did the *Lancet* article do?

	Chromosome	Gene(s) of interest within or near associated interval	Major allele, minor allele (minor allele frequency)*	Modelled allele	Effect of modelled allele on plasma HDL cholesterol (mmol/L)*	Effect of modelled allele on plasma triglycerides (mmol/L)*	Effect of modelled allele on plasma LDL cholesterol (mmol/L)*	Sample size (MI cases/ MI-free controls)	For modelled allele, observed change in MI risk (%; 95% CI)	For modelled allele, p value for association with MI
rs17482753	8p21	*LPL*†	G, T (0·10)	T	0·08	−0·24	..	19 139/50 812	−12% (−16 to −7)	4×10^{-7}†
rs17321515	8q24	*TRIB1*†	A, G (0·45)	G	0·02	−0·11	−0·05	19 139/50 812	−7% (−9 to −4)	2×10^{-6}†
rs6589566	11q23	*APOA1-APOC3-APOA4-APOA5*†	A, G (0·07)	A	0·05	−0·27	−0·09	18 310/49 897	−10% (−15 to −5)	8×10^{-5}†
rs4846914	1q42	*GALNT2*†	A, G (0·40)	A	0·02	−0·03	..	19 139/50 812	−3% (−6 to −1)	0·02†
rs2967605	19p13	*ANGPTL4*†	C, T (0·16)	C	0·05	−0·07	..	13 595/16 423	−5% (−10 to −1)	0·03†
rs3764261	16q13	*CETP*†	C, A (0·32)	A	0·10	..	−0·03	16 503/46 576	−4% (−7 to 0)	0·04†
rs61755018 (Asn396Ser)	18q21	*LIPG*	A, G (0·015)	G	0·14‡	..	..	17 165/49 077	−6% (−18 to 9)	0·41
rs17145738	7q11	*MLXIPL*	C, T (0·11)	T	0·03	−0·15	..	19 139/50 812	−1% (−4 to 3)	0·61
rs3890182	9q31	*ABCA1*	G, A (0·14)	G	0·03	..	0·05	19 139/50 812	−1% (−5 to 4)	0·76
rs2338104	12q24	*MMAB, MVK*	G, C (0·46)	G	0·03	..	..	19 139/50 812	0% (−3 to 3)	0·85
rs471364	9p22	*TTC39B*	T, C (0·12)	T	0·03	..	..	15 693/47 098	0% (−5 to 5)	0·97
rs2271293	16q22	*LCAT*	G, A (0·11)	A	0·03	..	..	19 139/50 812	4% (−1 to 8)	0·10
rs174547	11q12	*FADS1-FADS2-FADS3*	T, C (0·33)	T	0·03	−0·06	..	19 139/50 812	3% (−1 to 6)	0·11
rs1800588	15q22	*LIPC*	C, T (0·22)	T	0·05	0·07	..	17 917/49 514	4% (0 to 7)	0·04
rs16988929	20q13	*HNF4A*	C, T (0·01)	T	0·01	..	..	17 041/20 137	31% (12 to 54)	9×10^{-4}

Is this a death sentence for the HDL hypothesis?

Where did I enter the cycle

conjecture – **design** – experiment – **analysis** – conjecture – . . .

I heard about Mendelian randomisation in April 2017. I was immediately shocked by some basic mistakes that the researchers were making

- **Selection bias:** The same GWAS dataset is used to select instrumental variables and estimating γ, their effect on the risk exposure (HDL).
- **Ignoring measurement error:** People did not consider sampling **fluctuations** of $\hat{\gamma} = \mathrm{lm}(A \sim Z)$ and assumed $\hat{\gamma} = \gamma$.
- **Unrealistic assumptions** about direct effects: In my **exploratory data analysis**, there seems to be universal direct effects and occasional outliers.

So I worked out a statistical method to address these problems:

- Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, Dylan S. Small (2019+). Statistical inference in two-sample summary-data Mendelian randomisation using robust adjusted profile score. To appear in *Annals of Statistics*.

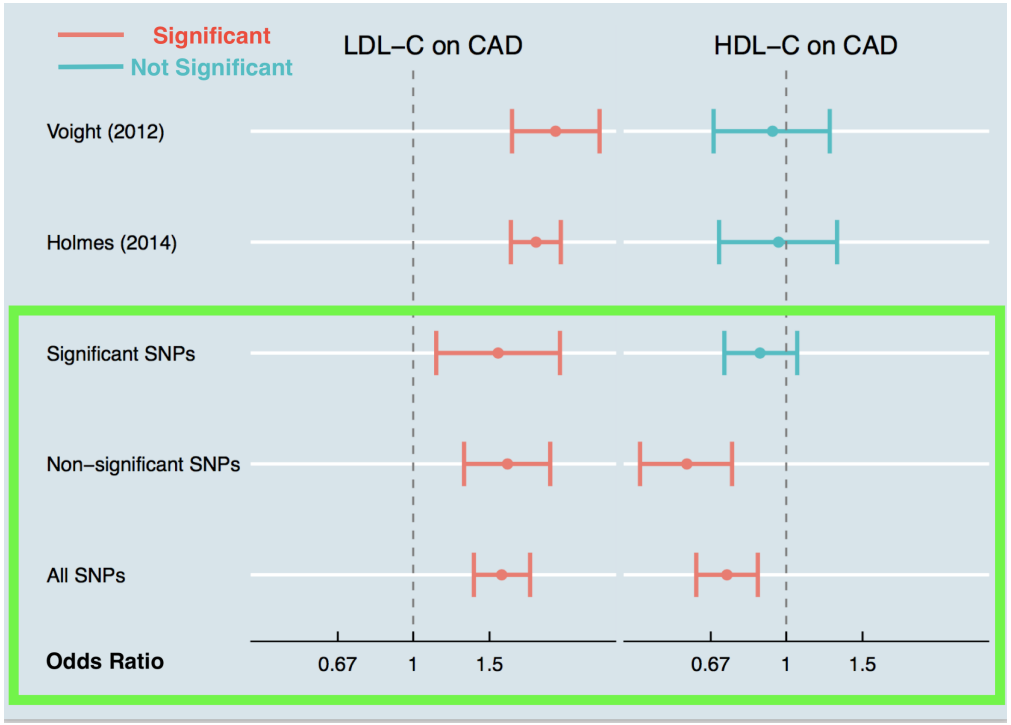
(I won't have time to go over the mathematical details in this talk, sorry!)

Where did I enter the cycle

I had good confidence in this method because

- All the assumptions were given careful considerations.
- The method seems to fit several dataset very well.

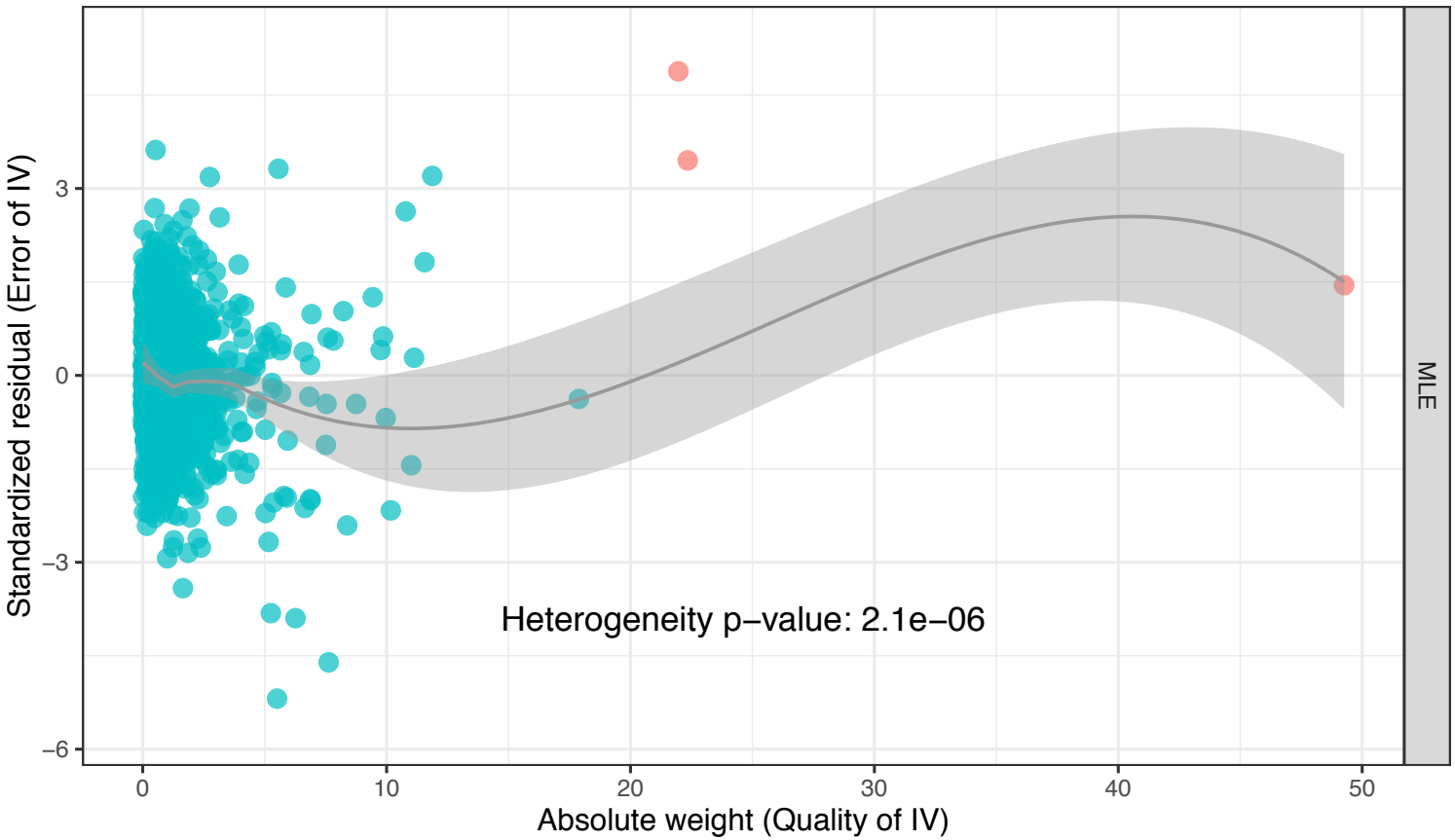
But when I applied it to HDL, something unexpected happened:



- Qingyuan Zhao, Yang Chen, Jingshu Wang, Dylan S. Small (2018). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomisation. To appear in *International Journal of Epidemiology*.

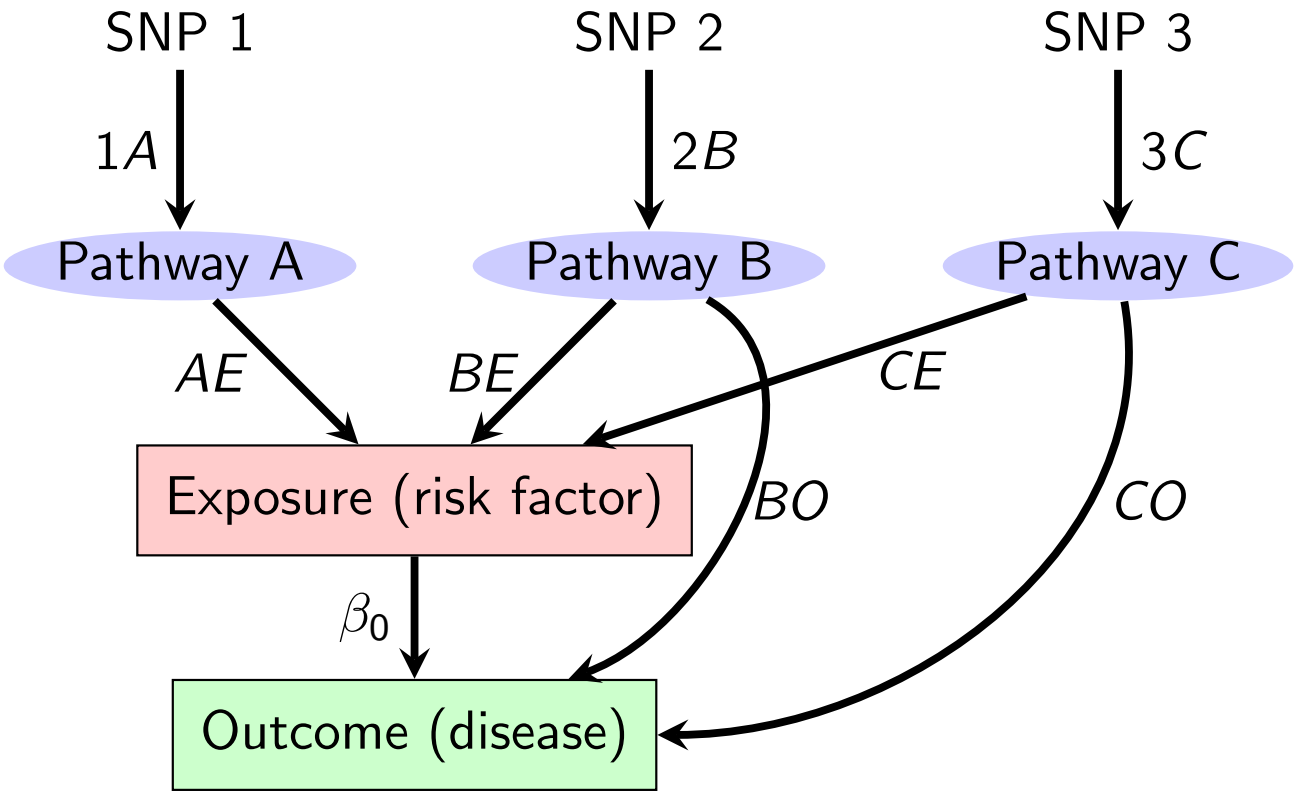
Heterogeneity among the instrumental variables

A diagnostic plot was developed to understand what happened.



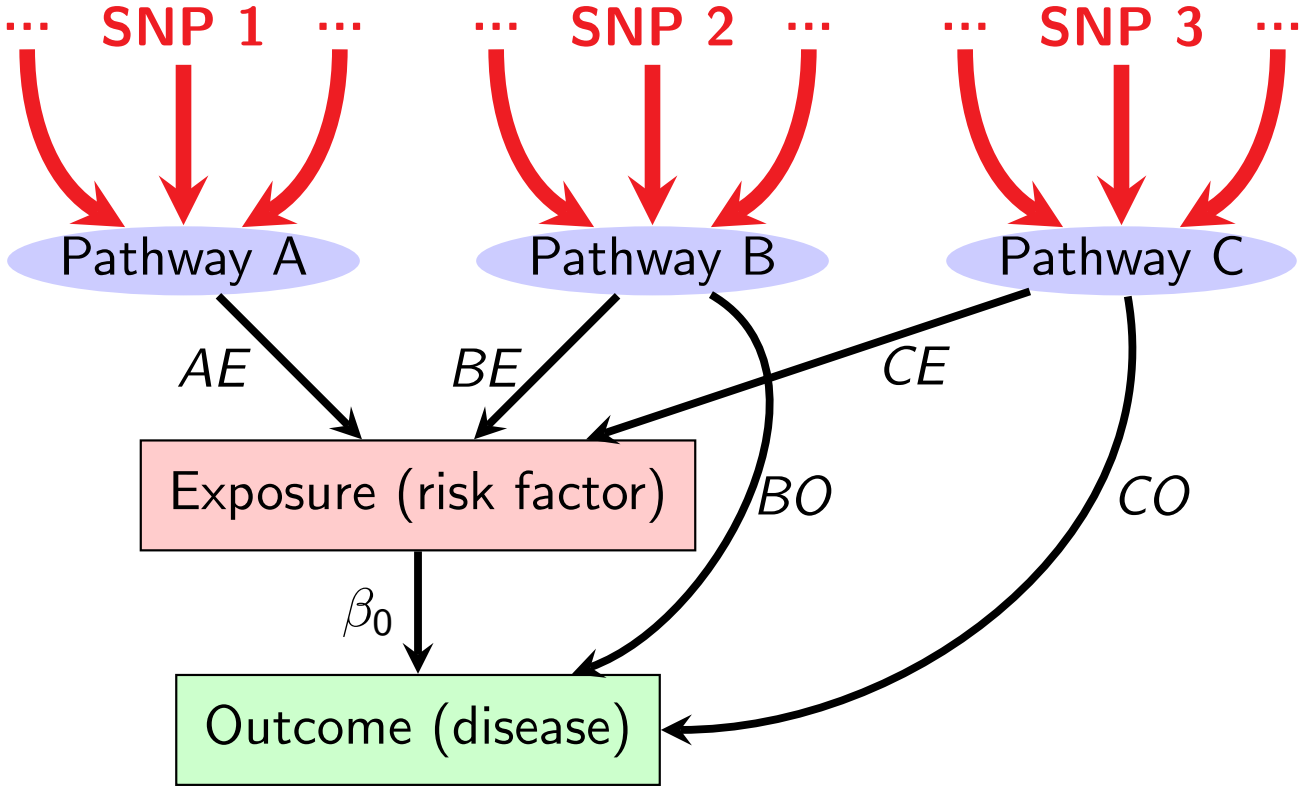
- x-axis is strengths of instrumental variables: $\hat{\gamma}$ divided by its standard error;
- y-axis is standardised residuals: $\hat{\Gamma} - \hat{\beta}\hat{\gamma}$ divided by its standard error.

Conjecture: Multiple pathways $\Longrightarrow$ multiple modes of β



	Exposure effect γ	Outcome effect Γ	**Ratio**
SNP 1	$1A \cdot AE$	$1A \cdot AE \cdot \beta_0$	β_0
SNP 2	$2B \cdot BE$	$2B \cdot BE \cdot \beta_0 + 2B \cdot BO$	$\beta_0 + (BO/BE)$
SNP 3	$3C \cdot CE$	$3C \cdot CE \cdot \beta_0 + 3C \cdot CO$	$\beta_0 + (CO/CE)$

Conjecture: Multiple pathways $\Longrightarrow$ multiple modes of β



	Exposure effect γ	Outcome effect Γ	**Ratio**
SNP 1	$1A \cdot AE$	$1A \cdot AE \cdot \beta_0$	$\boldsymbol{\beta_0}$
SNP 2	$2B \cdot BE$	$2B \cdot BE \cdot \beta_0 + 2B \cdot BO$	$\boldsymbol{\beta_0 + (BO/BE)}$
SNP 3	$3C \cdot CE$	$3C \cdot CE \cdot \beta_0 + 3C \cdot CO$	$\boldsymbol{\beta_0 + (CO/CE)}$

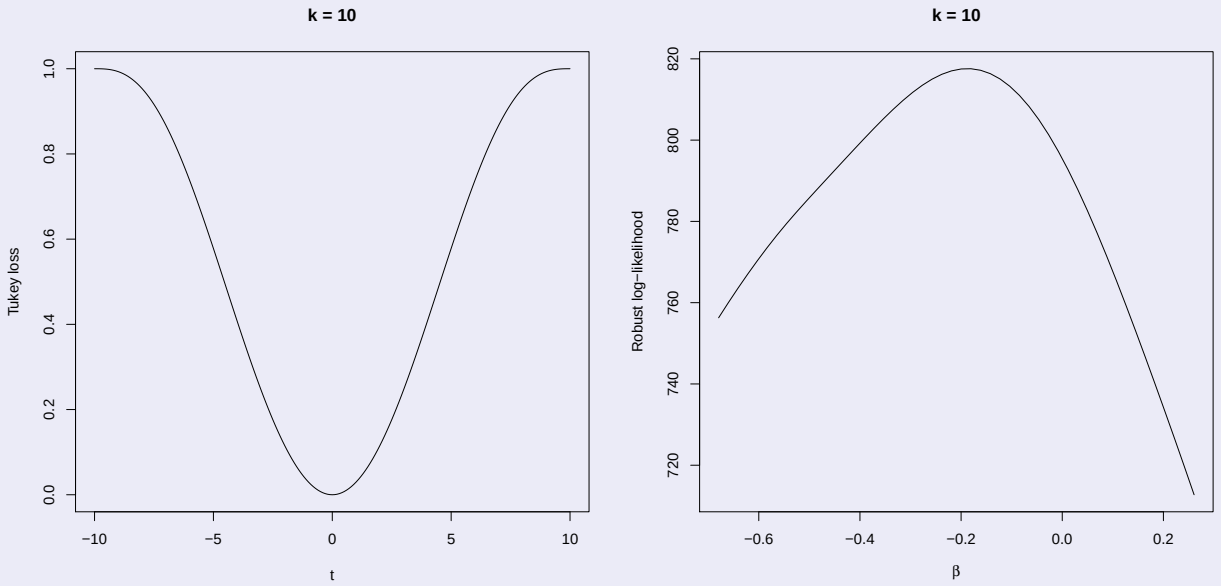
Detection via modal plot

- $l(\beta) = -\frac{1}{2}\sum_{j=1}^{p}\frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{1+\beta^2}$ penalises too much on "outliers".
- We can plot "robust" log-likelihood and search for **multiple modes**:

$$l_\rho(\beta) = -\sum_{j=1}^{p}\rho\left(\frac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{1+\beta^2}}\right).$$

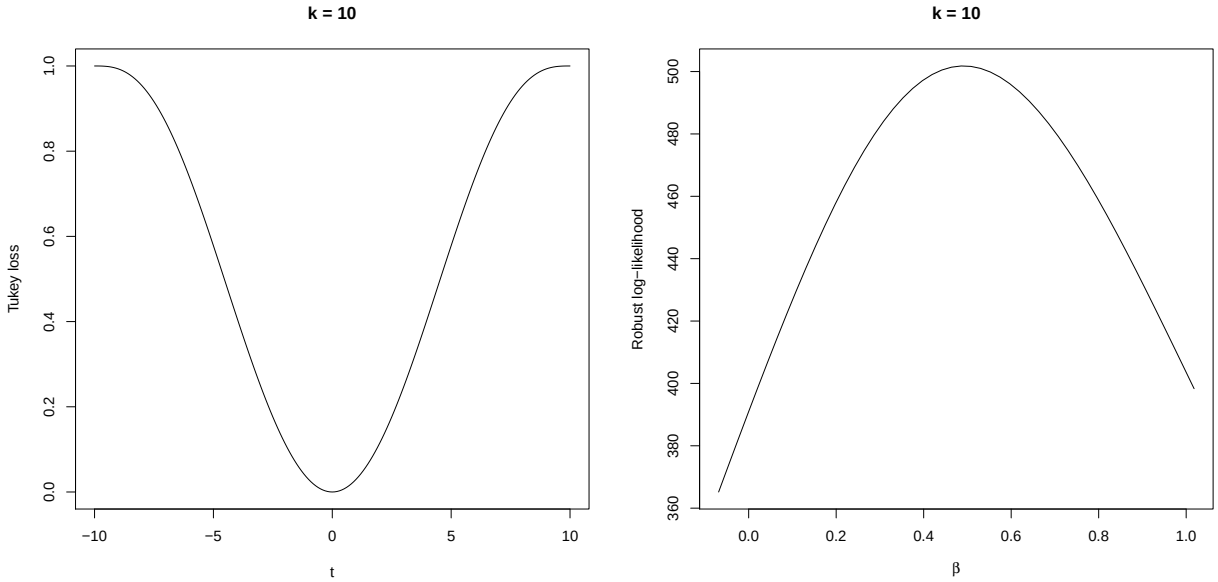
Example: Effect of HDL cholesterol on CAD

Left: loss function ρ; Right: robust log-likelihood.

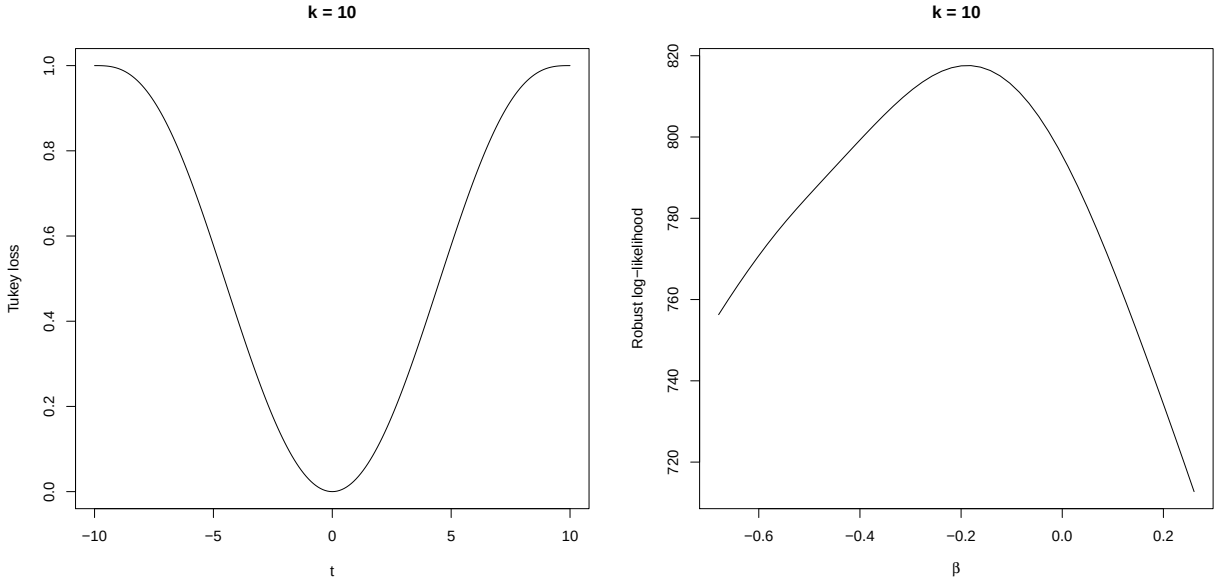


Compare with the modal plot for LDL-C

LDL-C

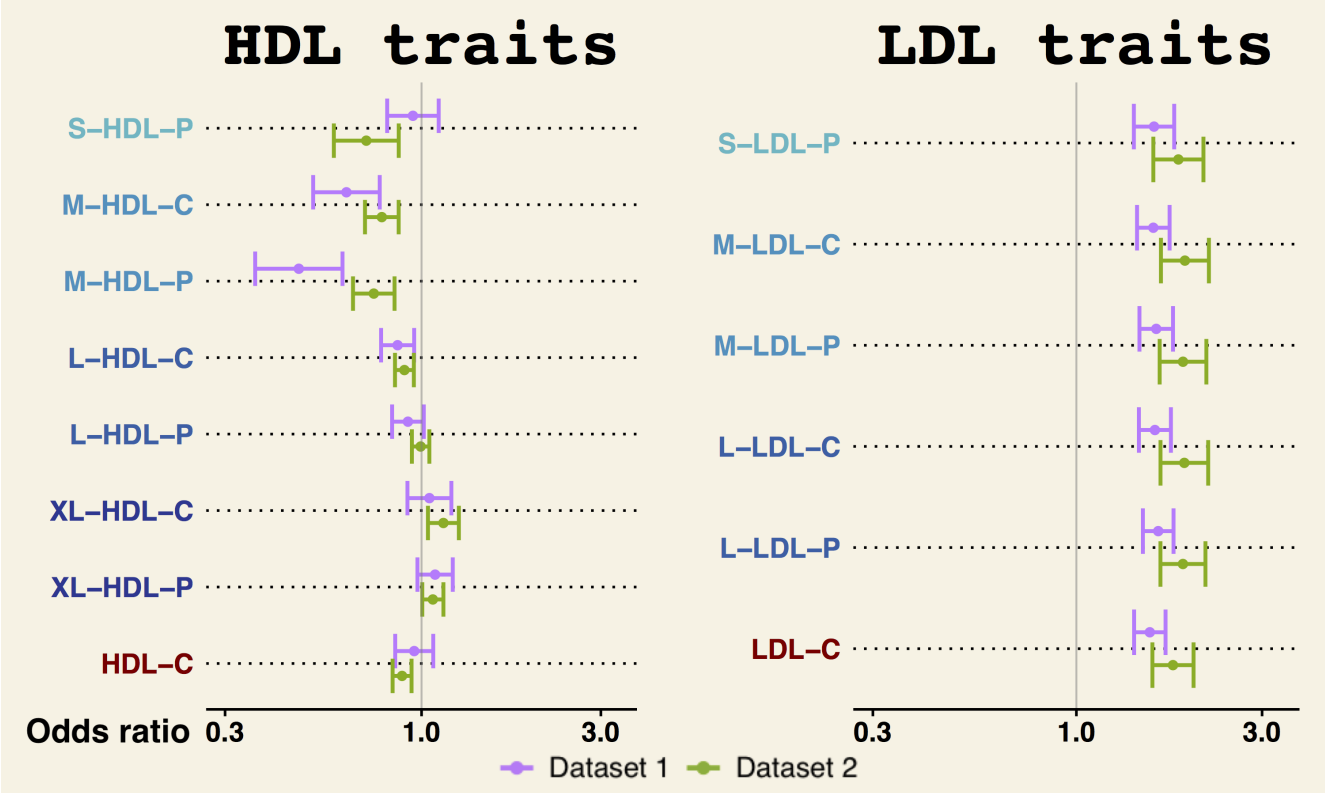


HDL-C



New analysis for lipoprotein subfractions

- If there are different pathways, HDL subfractions may show heterogeneous effects in Mendelian randomisation.
- A subsequent analysis was developed, and this conjecture is indeed true:



- This gives some support for the **HDL function hypothesis**.
- **The cycle of statistical research will continue**, until we fully understand the cardio-metabolic role of HDL particles.

Rest of the talk

1. Example 2: the epidemic growth of the COVID-2019 outbreak.

Timeline of the COVID-2019 outbreak

30 Dec. 2019	Health Commission in Wuhan, China announced 27 cases of viral pneumonia.
6 Jan. 2020	The causative pathogen was identified as a novel coronavirus (originally called 2019-nCoV, then COVID-2019).
20 Jan. 2020	An eminent Chinese epidemiologist first confirmed human-to-human transmission to the public in a televised interview.
23 Jan. 2020	Wuhan was put under quarantine: public transportation into/out of the city were halted, followed by stricter travel restriction within the city.
Mid Feb. 2020	621 confirmed cases among 3,700 passengers and crew on Diamond Princess. Governments in Japan, Singapore, and South Korea can no longer trace the epidemiological contact of many new cases.
19 Feb. 2020	More than 75,000 infected globally (about 60% are in Wuhan) and 2,000 deaths.

Where did I enter the cycle

conjecture – **design** – experiment – analysis – conjecture – . . .

- Wuhan is my hometown so I followed the news closely since the first announcement on 30 December, 2019.
- I saw the conclusions of two articles: "In its early stages, the epidemic doubled in size every 7.4 days." (Li et al, *NEJM*); "The epidemic doubling time was 6.4 days (95% CrI 5.8—7.1)" (Wu et al, *Lancet*).

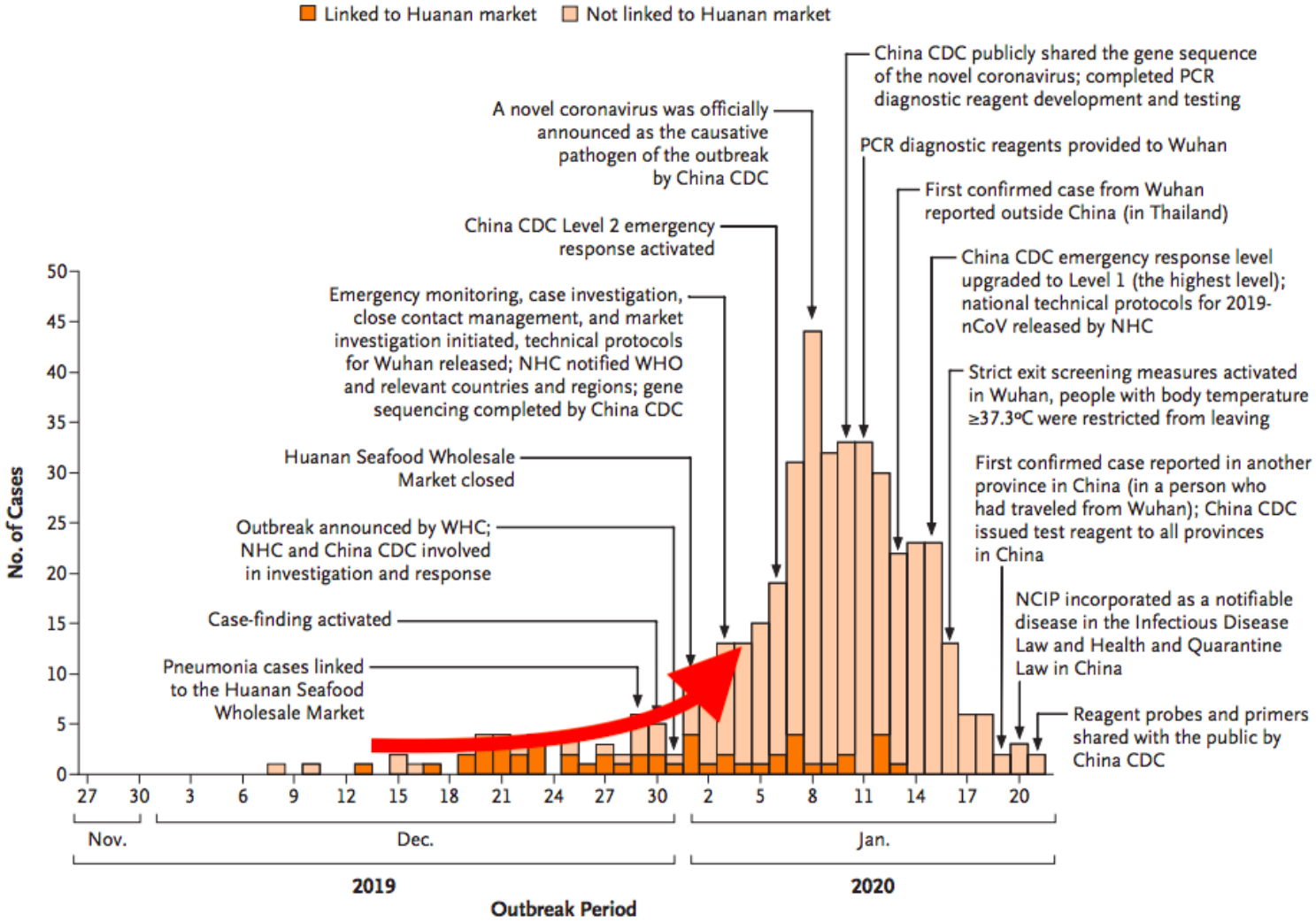
Why these numbers were impossibly low

- Suppose the epidemic starts on 1 December, 2019;
- Suppose the epidemic was doubling every 6.4 days.

Then we would have $2^{62/6.4} = 825$ people **infected** by 1 February, 2020. But we have a total of $14,380$ **confirmed** cases in China on 1 February, 2020.

- The numbers still don't add up if we consider zootonic exposure (animal-to-human transmission).

What did the *NEJM* paper do?



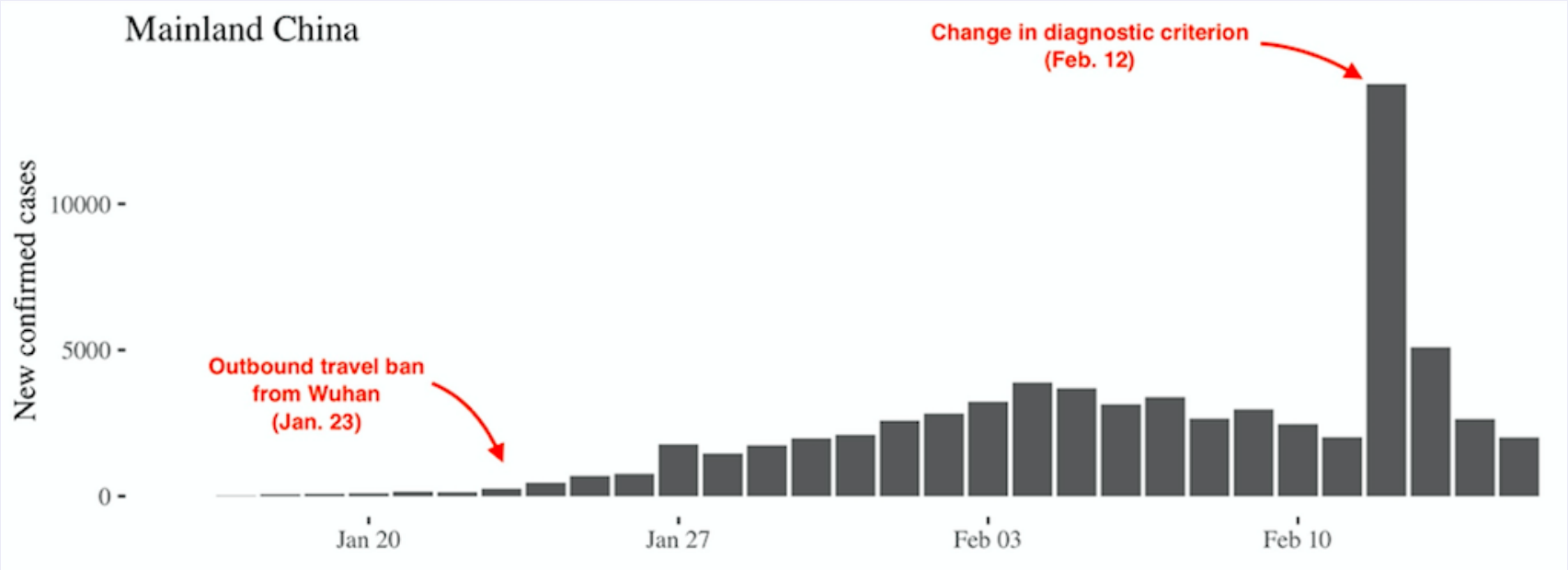
- Data: Symptom onset of first 425 confirmed cases in Wuhan.
- Model: Exponential growth of case counts up to 4 January, 2020.

The obvious challenge

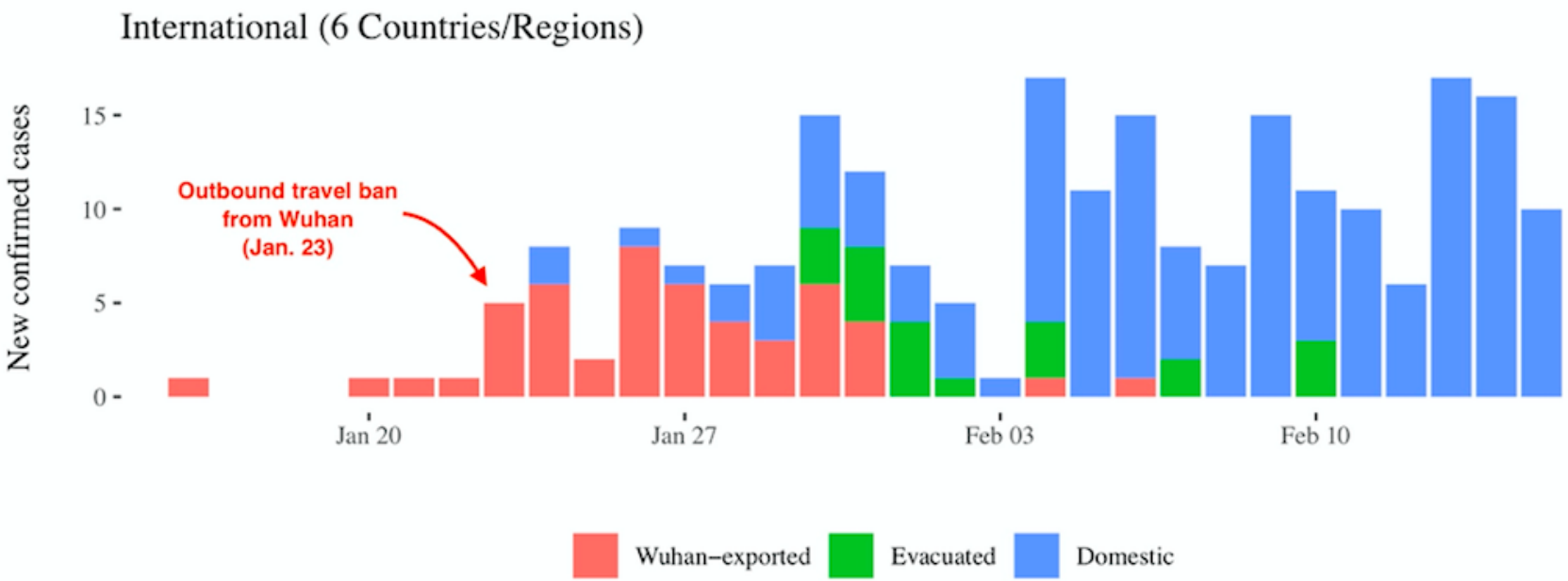
Data in Wuhan are biased because of the very strict diagnostic criterion. (To be fair, this is why the *NEJM* paper only used symptom onsets up to 4 January, 2020, but seriously?)

Change in diagnostic criterion

- **15 January:** Only cases with direct exposure to Huanan seafood market meet the diagnostic criterion.
- **12 February:** Positive results using RT-PCR array no longer required.



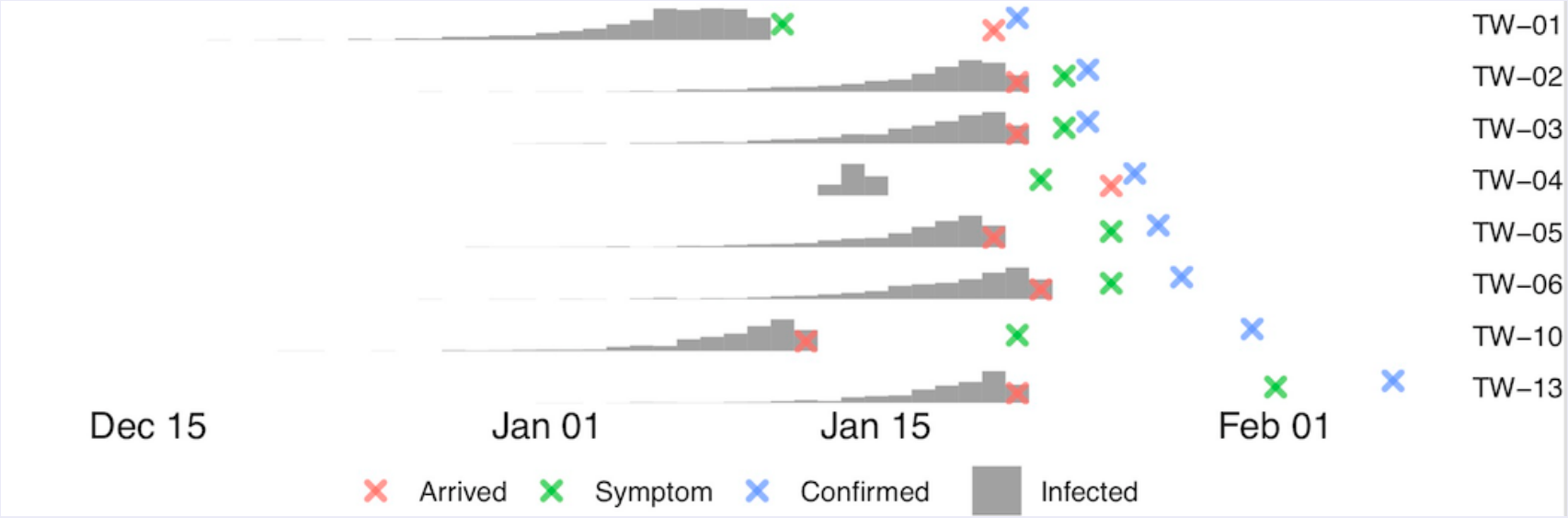
Key idea 1: Use international data



- A total of 50 cases in Hong Kong, Japan, South Korea, Macau, Singapore, Taiwan exported from Wuhan.
- (Hopefully) Free from selection bias due to delayed diagnosis.

Key idea 2: Simulate infection time

Dataset



Simulate infection time

Infected = Symptom - **Incubation period**, **truncated by travel history**.

- We used previously reported **incubation period** (mean = 5.2 days).

Advantages of using infection time:

1. Travel history allows us to narrow down the infection time.
2. We can directly account for the 23 January travel ban.

Key idea 3: Model the 23 January quarantine

A simple model

- Let WI_t be the number of infected people in Wuhan on day t. We assume it was growing exponentially:

$$WI_t = WI_0 \cdot e^{rt},\ 0 \le t \le T.$$

- T corresponds to 23 January, 2020.
- We further assume a small fraction OR of the Wuhan population traveled to the international destinations every day, before outbound travel was banned on 23 January:

$$OI_t \sim \text{Poisson}(\lambda_t),\ \lambda_t = WI_t \cdot \left(1 - \prod_{s=t}^{T}(1-OR)\right) \approx (WI_0 \cdot OR) \cdot e^{rt}(T-t+1).$$

- The adjustment term $(T-t+1)$ is important. It predicts that λ_t has a stationary point at $t = N + 1 - 1/r$.

Statistical inference

Point estimator

- Can estimate λ_t by counting the simulated infection time falling on day t.
- Our model of λ_t says

$$\log(\lambda_t) - \log(T - t + 1) = r \cdot t + \log(WI_0 \cdot OR).$$

- Can estimate r by linear regression $\log(\lambda_t)$ for with an offset $\log(T - t + 1)$.

Bayesian inference

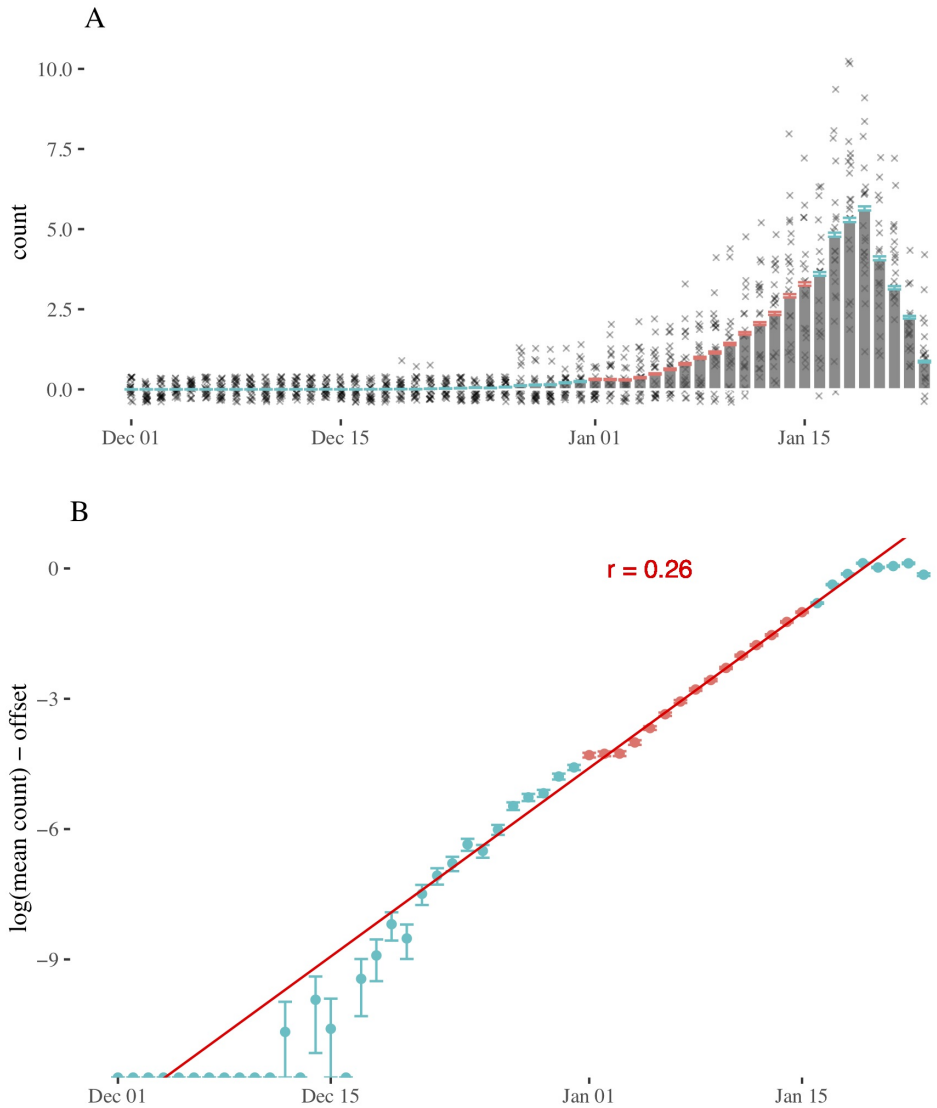
- The point estimator ignores the "fluctuations" in estimating λ_t.
- Bayesian posterior:

$$\pi(r \mid \text{Data}) = \int \pi(r \mid OI) \cdot \mathbb{P}(OI \mid \text{Data}) \, \mathrm{d}OI,$$

$$\pi(r \mid OI) \propto \pi(r) \, \mathbb{P}(OI \mid r).$$

- Put diffuse prior for r (and OR and WI_0).

Point estimator



- $r = 0.26$ corresponds to stationary point of λ_t at 20 January, 2020.

Results of Bayesian analysis

		Incubation period	
		Mean = 5.2 days Std. deviation = 3.7 days (Li et al., 2020)	Mean = 6.5 days Std. deviation = 2.6 days (Wu et al., 2020)
Using infections on January 1---20	Growth exponent	**0.25 (0.17 – 0.34)**	0.25 (0.18 – 0.32)
	Doubling time (days)	**2.8 (2.0 – 4.0)**	2.9 (2.2 – 4.0)
	Basic reproduction number	**5.7 (3.4 – 9.3)**	5.5 (3.5 – 8.3)
Using infections on January 1---23	Growth exponent	0.22 (0.16 – 0.29)	0.21 (0.15 – 0.27)
	Doubling time (days)	3.2 (2.4 – 4.3)	3.4 (2.6 – 4.5)
	Basic reproduction number	4.8 (3.2 – 7.0)	4.4 (3.0 – 6.2)

- Estimated growth was much faster than initial reports.

Comparison with the *Lancet* article

- The *Lancet* article reported a doubling time of 6.4 days (95% CrI 5.8—7.1).
- They also used international cases (up to 25 January, 2019).
- They used standard (and much more complicated) SEIR model for epidemics:

$$\frac{dS(t)}{dt} = -\frac{S(t)}{N}\left(\frac{R_0}{D_I} I(t) + z(t)\right) + L_{I,W} + L_{C,W}(t) - \left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N}\right) S(t)$$

$$\frac{dE}{dt} = \frac{S(t)}{N}\left(\frac{R_0}{D_I} I(t) + z(t)\right) - \frac{E(t)}{D_E} - \left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N}\right) E(t)$$

$$\frac{dI(t)}{dt} = \frac{E(t)}{D_E} - \frac{I(t)}{D_I} - \left(\frac{L_{W,I}}{N} + \frac{L_{W,C}(t)}{N}\right) I(t)$$

Comparison with the *Lancet* article

What did they miss?

They did not consider the 23 January quarantine!

- If we did not include the $T - t + 1$ term in our model, our estimate would be as low as theirs.

How did we not miss this term?

We thought about how the data were generated. Any Wuhan-exported patient went through the following steps:

Arrived in Wuhan $\rightarrow$ Infected $\rightarrow$ Left Wuhan $\rightarrow$ Confirmed a COVID-2019 case.

- The patient could show symptom **before or after** they left Wuhan!
- The *Lancet* paper only models the symptom onset, so it did not occur to them that the 23 January quarantine needs to be considered.

Take-home messages

Cycle of statistical research

- Every statistical researcher needs to take the cyclic view.
- Understanding the principles in causal inference can be helpful: it forces us to think about the underlying **data generating mechanism** and **how to collect data**.

More about causal inference

- "New" Part III course in the Michaelmas term (`http://www.statslab.cam.ac.uk/~qz280/teaching/Causal_Inference_2019.html`).
- "New" reading group (`http://talks.cam.ac.uk/show/index/105688`).