The Institute of Actuaries of Japan
Think the Future, Manage the Risk

Life and Health Risk Assessment Using Non-Traditional Data with Causal Inference

Kazunori Nishio

FIAJ, Sumitomo Life Insurance Company

IAA Life Section Webinar
2024-10-10

Background

The use of non-traditional data has been increasing in life insurance companies.
Interest in causal relationships related to life and health risks is growing.

Today's Topic

Discuss how actuaries can incorporate causal inference into risk assessment.

Causal Inference

Basic Concept → **Major Method** → **Latest Method**

Agenda

I. Thought Process in Causal Inference

II. Causal Inference Methods

- Propensity Score Matching
- Instrumental Variables Method
- Methods Using Machine Learning

Thought Process in Causal Inference

Causal problems fundamentally differ from prediction problems.

Prediction problem

We use past data to predict future outcomes.

- How likely is the customer to purchase the product?

Causal problem

We consider what happens when there is a change.

- How *much more* likely is the customer to purchase the product if we recommend it?

Thought Process in Causal Inference

The effect of recommendations is verified through A/B testing in marketing.

Causal Problems in Risk Assessment

A/B testing is challenging for causal problems in risk assessment.
Using natural experiment data requires the introduction of rigorous theories.

Prediction problem

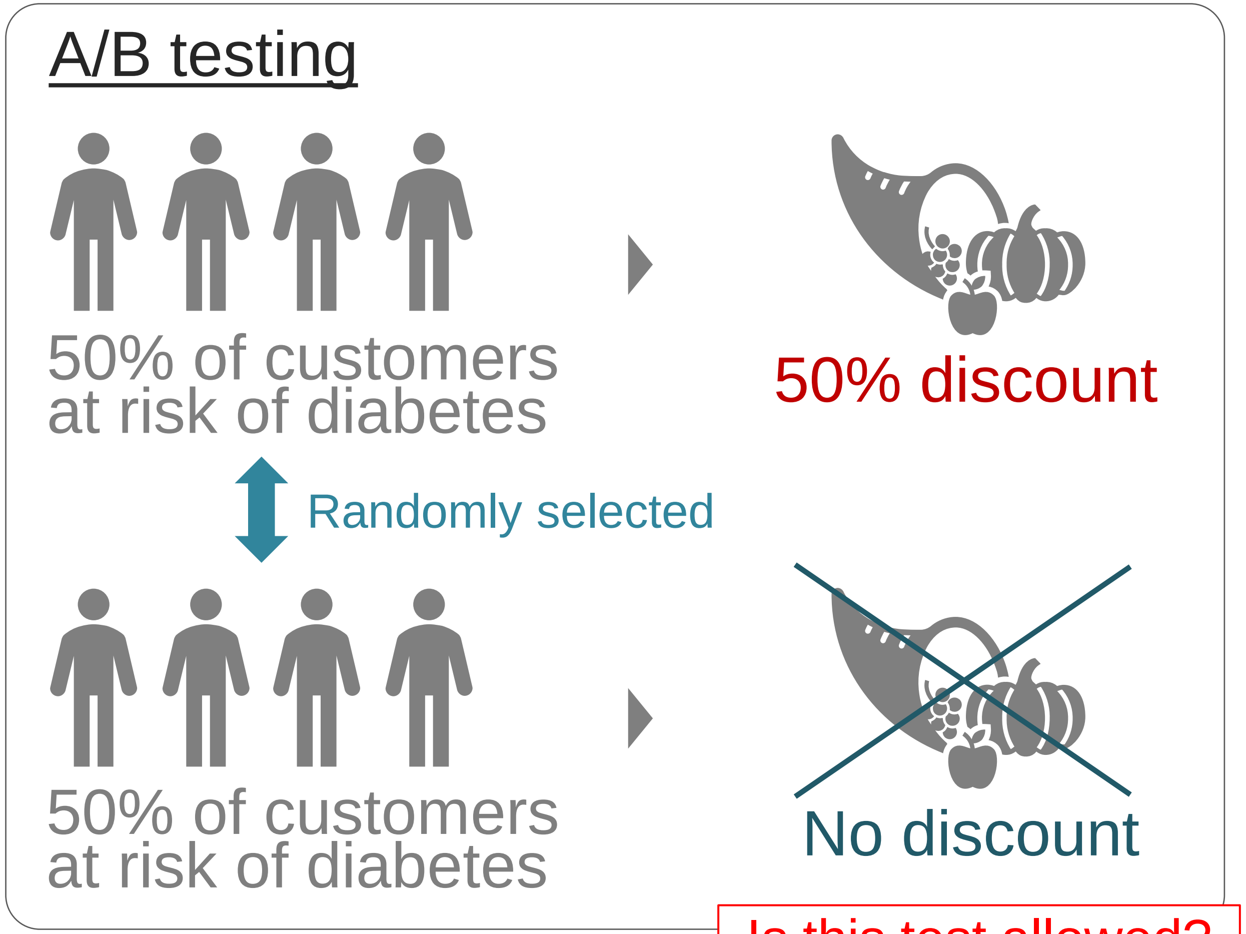
We use past data to predict future outcomes.

- How likely is the customer to develop diabetes?

Causal problem

We consider what happens when there is a change.

- How *much less* likely is the customer to develop diabetes if they receive a 50% discount on a healthy food purchase?



Is this test allowed?

Causation and Correlation

Understanding the difference between causation and correlation is the first step.

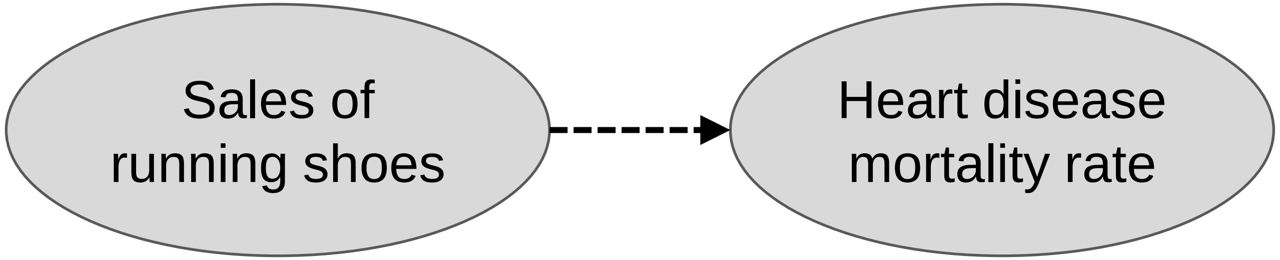
Does the sales of running shoes cause a decrease in heart disease mortality rates?

Correlation

A relationship between two variables that are associated with each other.

Causation

A relationship where one event causes another.



Causation?

Reference

American Academy of Actuaries (2022), *An actuarial view of correlation and causation – from interpretation to practice to implications.*

Causation and Correlation

The unobserved variable "health consciousness" can cause spurious correlations. Identifying and removing such confounders is critical in causal inference.

Correlation

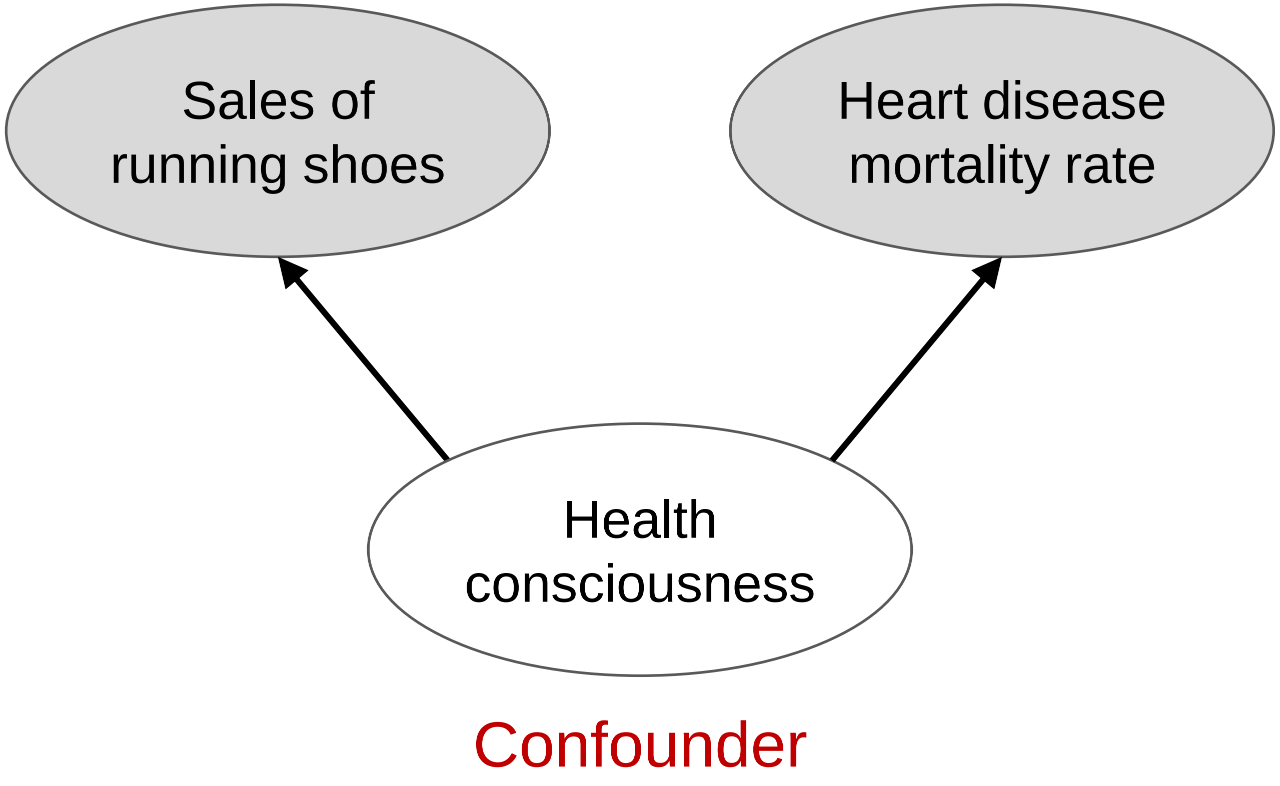
A relationship between two variables that are associated with each other.

Causation

A relationship where one event causes another.

Numerical difference

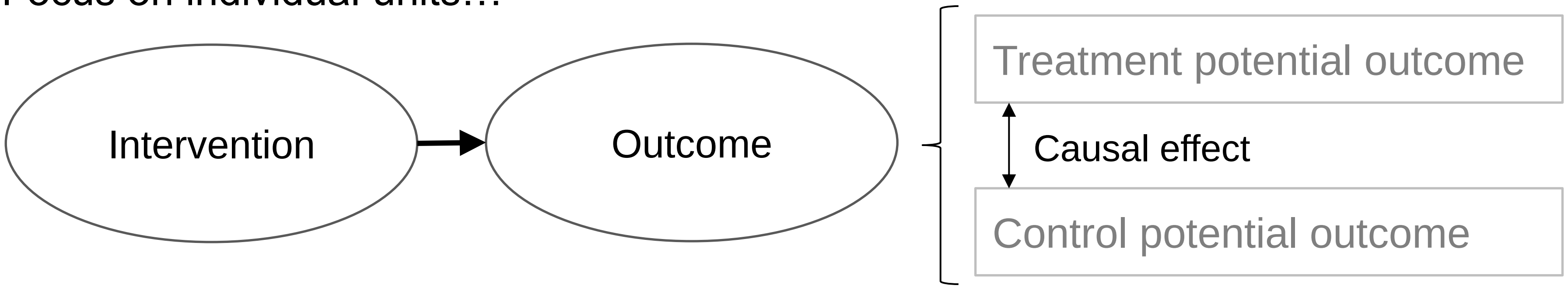
The numerical difference between causation and correlation is introduced by bias.



Concept of Causal Inference

Causal effects are estimated using the potential outcomes framework.
Estimating these effects becomes challenging when using natural experiment data.

Focus on individual units…

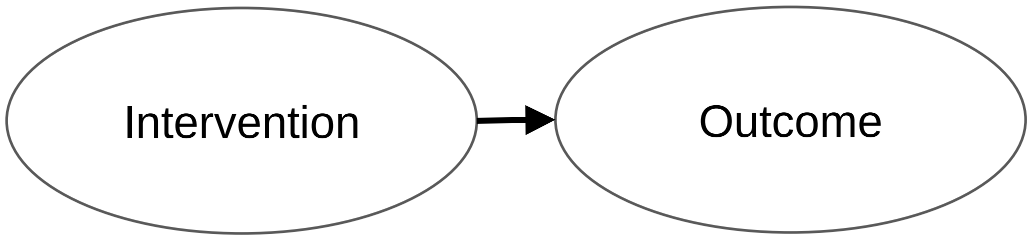


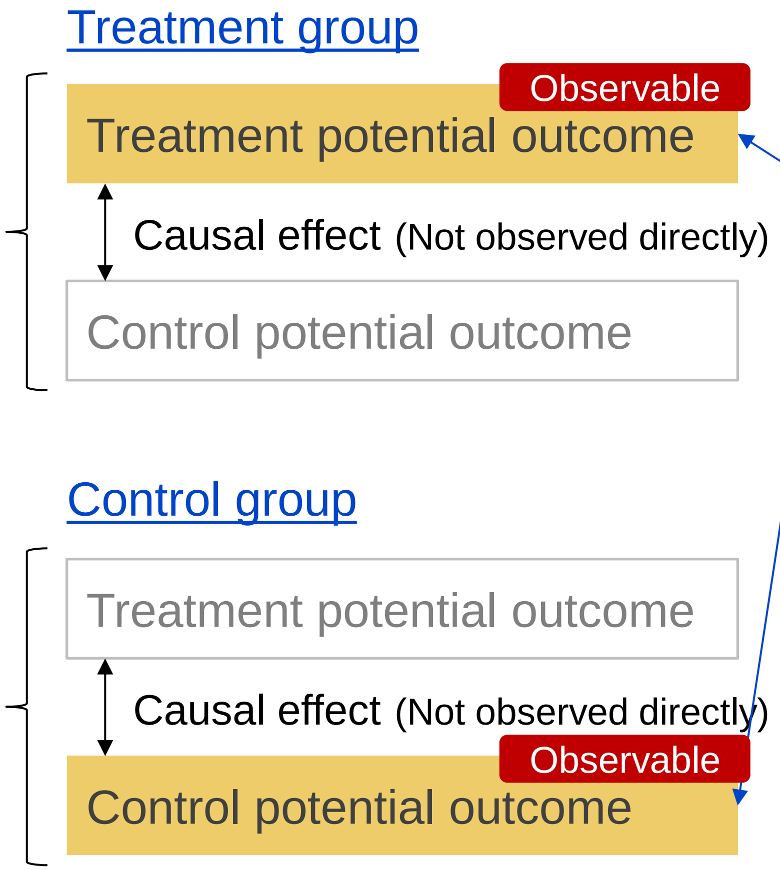
In actual observation, we can only observe the treatment potential outcome or the control potential outcome for one individual unit.

Concept of Causal Inference

Causal effects are estimated using the potential outcomes framework.
Estimating these effects becomes challenging when using natural experiment data.

Focus on a group of individual units…





Average Treatment Effect (Estimate of causal effect)

- Difference between comparable groups
- In A/B testing, this calculation is straightforward.
- In the analysis of observational data, simple comparisons introduce bias due to confounders.

Removing Confounding Bias

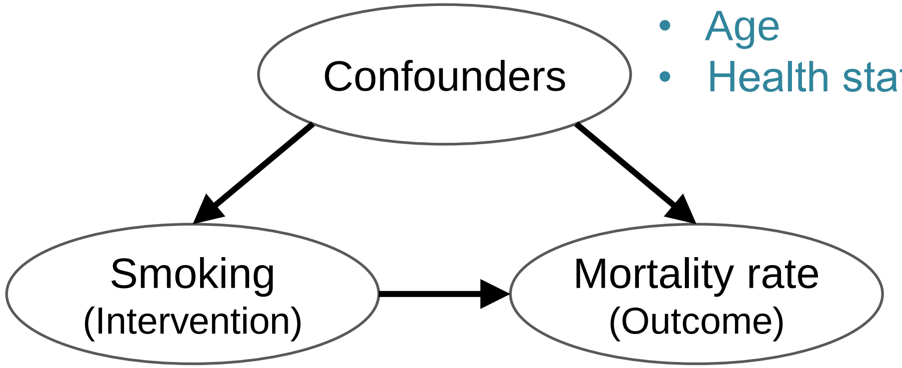
There are two fundamental ideas for removing confounding bias.
Causal inference theory seeks to minimize bias based on these principles.

Identify and observe all confounders

This requires domain knowledge, even when using approaches such as causal discovery.
Collecting data on confounders is equally important.

Align confounders between the treatment group and the control group

Ensure comparability across the two groups; otherwise, bias may be introduced.



- Gender
- Age
- Health status …

Evidence Levels of Causal Methods

A realistic goal is to understand how close we are to understanding the causality.
In risk assessment, judgements on causality must be made based on the situations.

High Level of Evidence (near causation)

Method	Description	Risk Assessment
Meta-analyses	Combining results from multiple studies or experiments	• Using external results
Experiments (A/B testing)	Dividing subjects into two groups and applying an intervention to one group	• Using external results • Unlikely within a company
Natural experiments	Leveraging the aforementioned experimental-like situations	• Using external results • Possibly feasible within a company
Regression analyses	Analyzing data at hand	• Possible internally, but low evidence level

Low Level of Evidence (near correlation)

Agenda

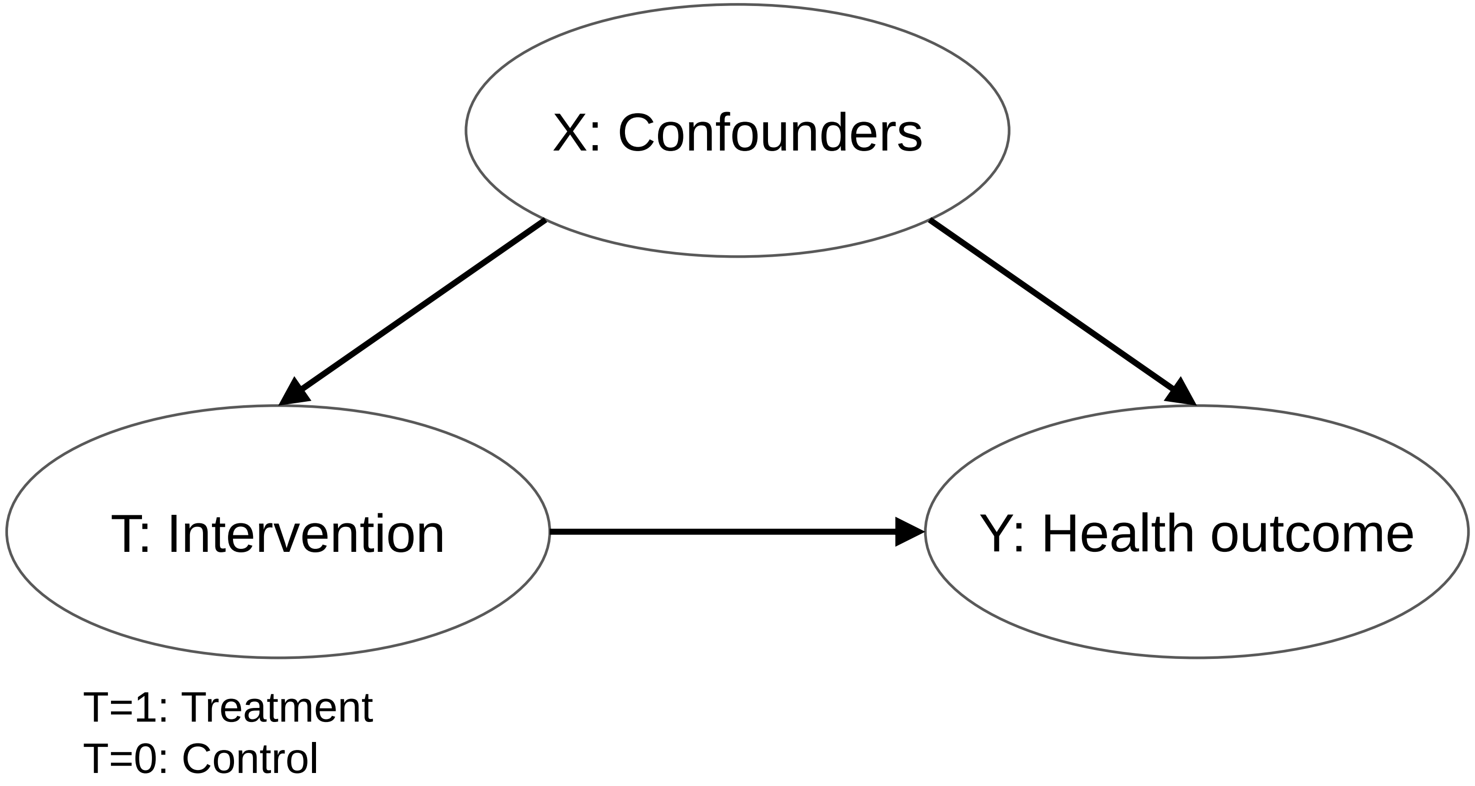
I. Thought Process in Causal Inference

II. Causal Inference Methods

- Propensity Score Matching
- Instrumental Variables Method
- Methods Using Machine Learning

Causal Inference Methods

Discuss specific methods for estimating causal effects in risk assessment.
Assume that the causal relationship is modeled and we have observational data.



Causal Inference Assumptions

The followings are overviews of some required assumptions.
Demonstrating that these are met is challenging and requires judgement.

Purpopse

Exchangeability (Ignorability)

The treatment assignment is independent of the potential outcomes, given a set of observed covariates.
(In general, exchangeability can be reduced to unconfoundedness.)

This is for accurately estimating the effect of the intervention on the outcome.

Positivity

Every subject has a positive probability of receiving each treatment.

This is to enable the estimation of causal effects for any individual.

SUTVA (Stable Unit Treatment Value Assumption)

The potential outcomes for any unit are unaffected by the interventions assigned to other units.

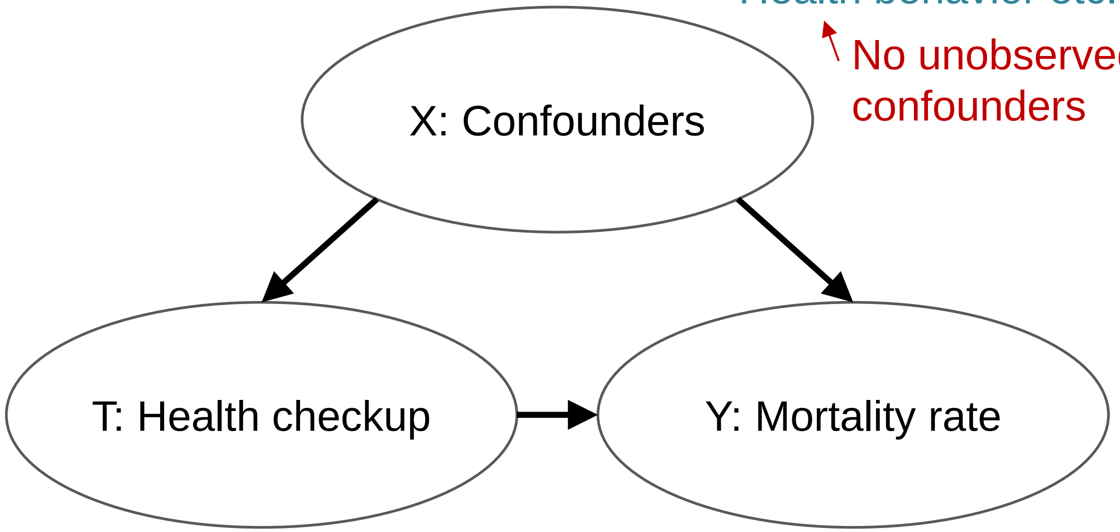
This is for independently evaluating the effect of the intervention.

Propensity Score Matching

Matches the treatment group and the control group using propensity scores.
This adjusts for the influence of confounders in causal effect estimation.

- Demographics
- Health status
- Health behavior etc.

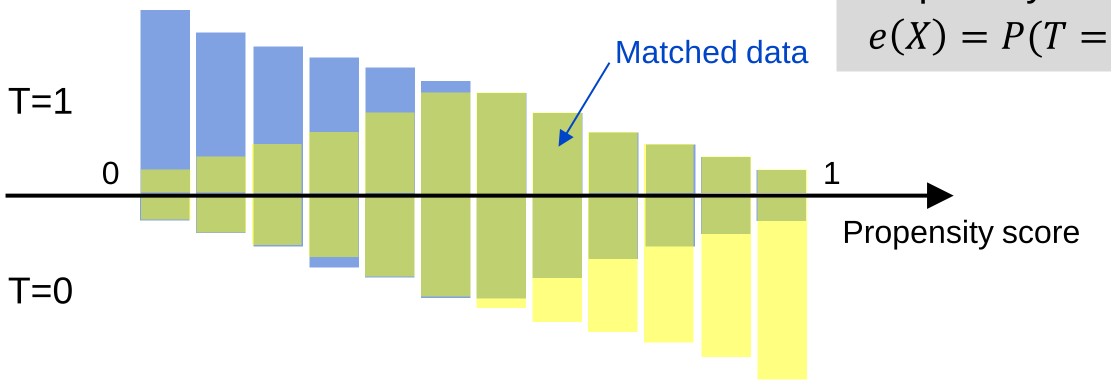
No unobserved confounders



T=1: Health checkup
T=0: No health checkup

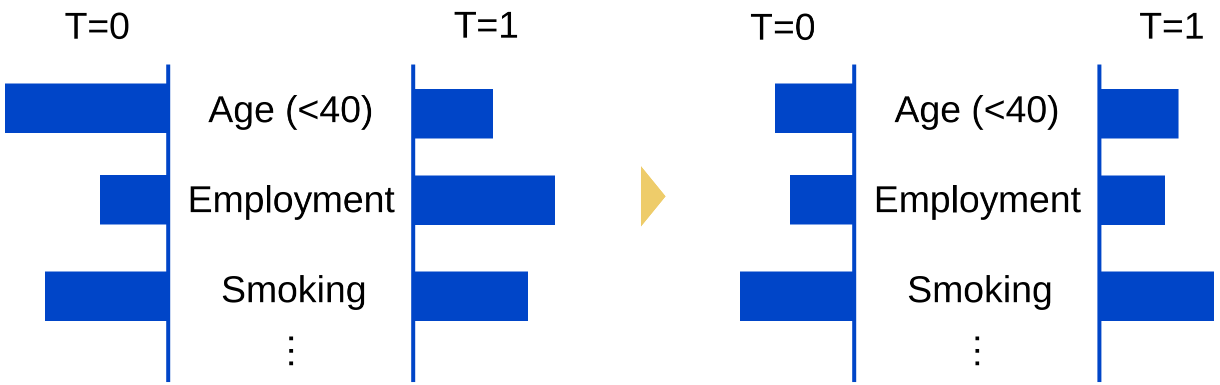
Propensity score matching

Propensity score:
$e(X) = P(T = 1|X)$



Before matching

After matching



Instrumental Variables Method

This method is applicable when there are unobserved confounders.

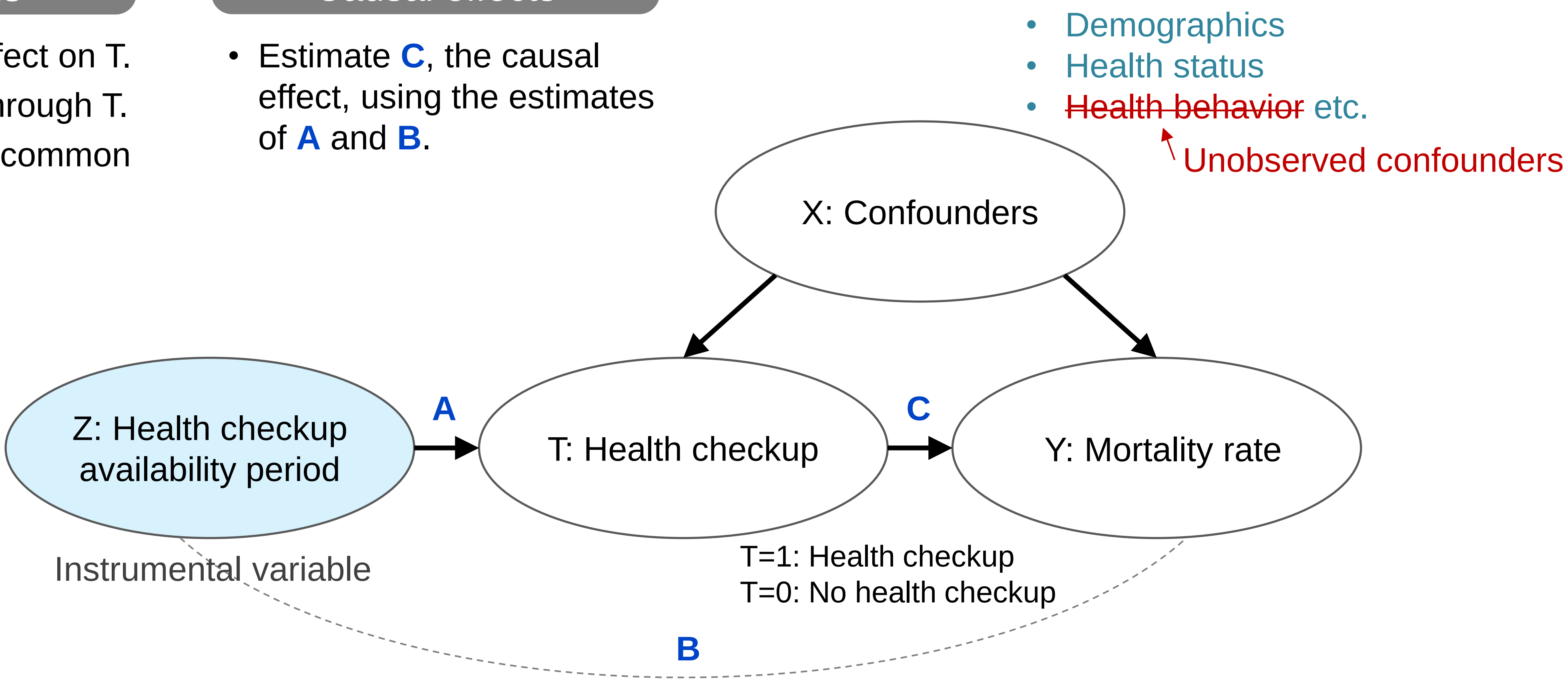
An instrumental variable affects the intervention but not the outcome directly.

Assumptions

- Z has a causal effect on T.
- Z affects Y only through T.
- Z does not share common causes with Y.

Causal effects

- Estimate **C**, the causal effect, using the estimates of **A** and **B**.



Methods Using Machine Learning

This allows for the estimation of Conditional Average Treatment Effects (CATE).
These represent causal effects for specific combination of attributes.

Average Treatment Effect

Age	Employment	Smoking	…	Causal effect
28	Yes	No		Estimate the group average
37	No	Yes		
64	No	Yes		
49	Yes	No		
55	Yes	No		
⋮	⋮	⋮		⋮

Health checkup case

In survey X, health checkups reduce mortality rates by p% for the entire group.

Conditional Average Treatment Effect

Age	Employment	Smoking	…	Causal effect
28	Yes	No		Estimate
37	No	Yes		Estimate
64	No	Yes		Estimate
49	Yes	No		Estimate
55	Yes	No		Estimate
⋮	⋮	⋮		⋮

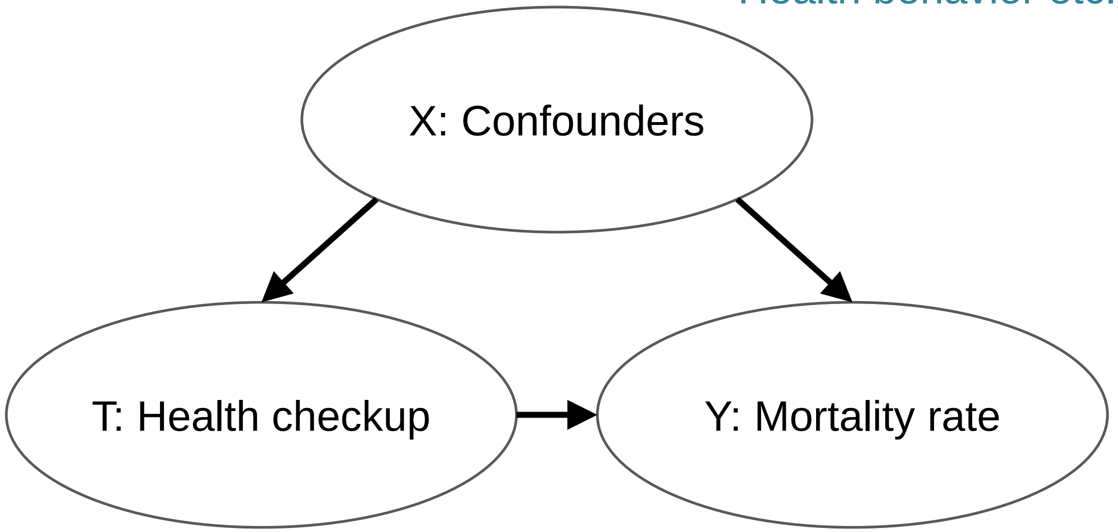
Health checkup case

In survey X, health checkups reduce mortality rates by q% for employed, non-smoking individuals in their 40s.

Meta-Learner

A meta-learner combines multiple machine learning models.
This integrates different trained models to estimate causal effects.

- Demographics
- Health status
- Health behavior etc.

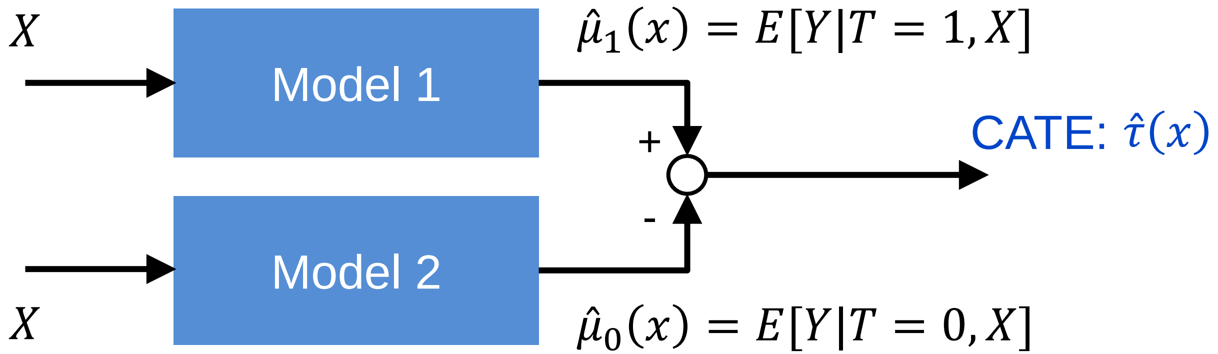


T=1: Health checkup
T=0: No health checkup

Meta-learner

S-learner	T-learner	X-learner
R-learner	DR-learner	The best model depends on the specific problem.

Example: T-learner

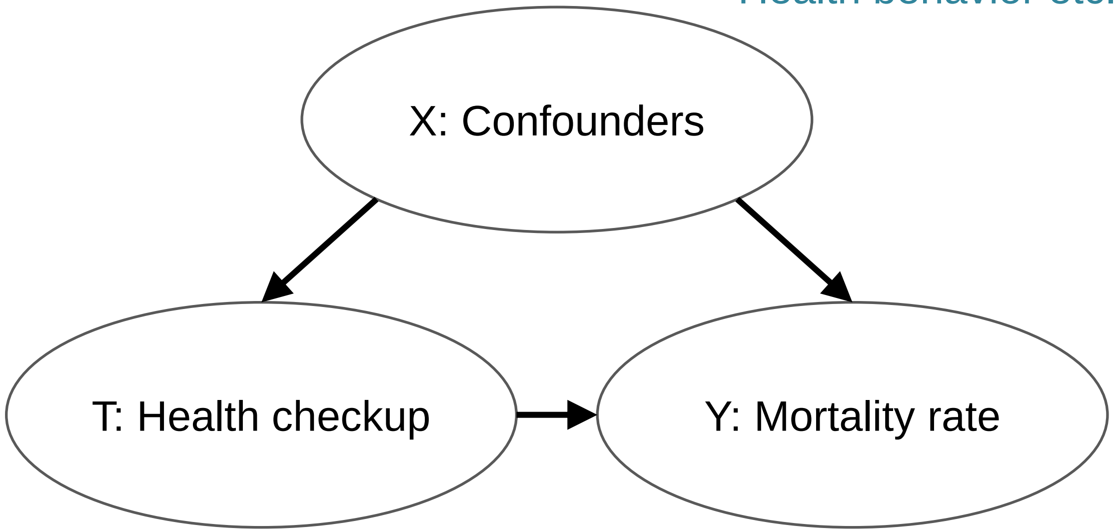


Causal Forest

Uses the random forest to estimate causal effects on individual data points.

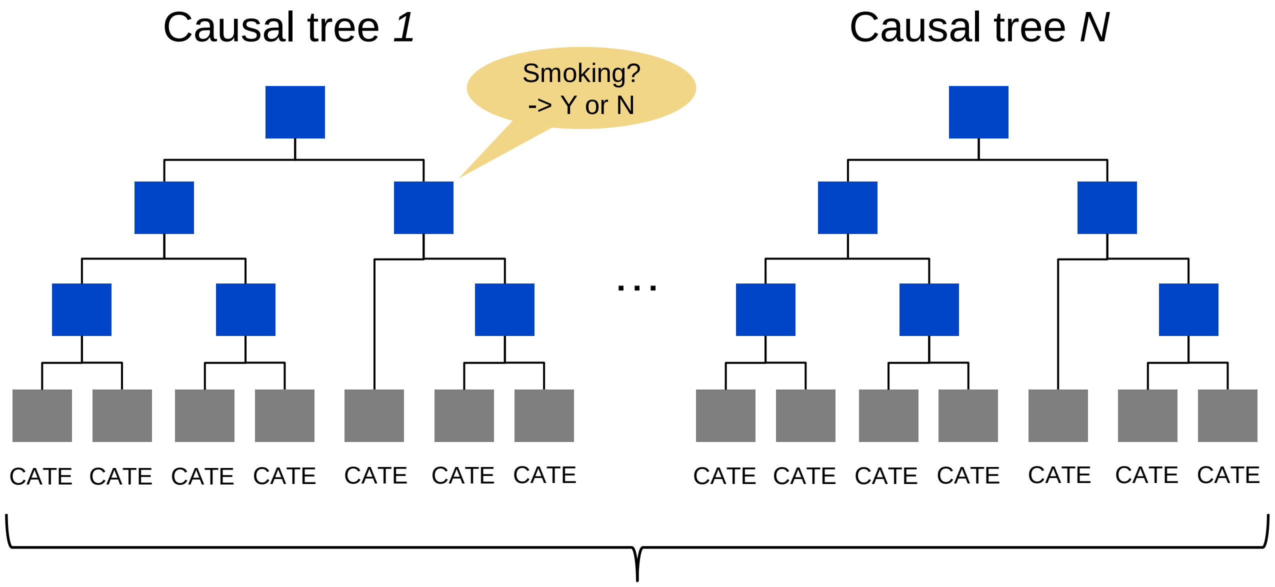
Highly capable of capturing non-linearities and complex interactions in the data.

- Demographics
- Health status
- Health behavior etc.



T=1: Health checkup
T=0: No health checkup

Model structure



Causal forest

CATE on each data point: $\hat{\tau}(x)$

Summary

We covered causal inference with a focus on its application to risk assessment.

- Thought Process in Causal Inference
- Propensity Score Matching
- Instrumental Variable Method
- Method Using Machine Learning

The increasing variety of data is likely to lead us to encounter causal problems more frequently.

I hope today's introduction will inspire interest in this field among actuaries.