A Gentle Introduction to Causal Mediation Analysis

Justin Armanini[1, 2, *]

December 17, 2025

[1]Models and Algorithms for Data and Text Mining (MADLab), University of Milano-Bicocca, Milan, Italy

[2]Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

[*]Corresponding author: j.armanini@campus.unimib.it

About me

Justin Armanini is a PhD candidate at the University of Milano-Bicocca, funded by Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy.

His research focuses on **merging Causality and Natural Language Processing (NLP)** methods to support clinical decisions.

Table of Contents

Causal Mediation Analysis theory

What is (Causal) Mediation Analysis?

- Mediation analysis is the study of how a treatment X influences an outcome Y **through one or more intermediate variables**, called mediators M
- Mediation analysis enables us to understand **the pathways by which a cause produces an effect**
- Causal mediation analysis performs such analysis through the lenses of **Causal Inference framework**
- While traditional mediation is prevalent in the literature, causal mediation has been gaining in popularity since the early 2000s as an alternative method for assessing mediation [1]

What is (Causal) Mediation Analysis?

- Mediation analysis is the study of how a treatment X influences an outcome Y **through one or more intermediate variables**, called mediators M
- Mediation analysis enables us to understand **the pathways by which a cause produces an effect**
- Causal mediation analysis performs such analysis through the lenses of **Causal Inference framework**
- While traditional mediation is prevalent in the literature, causal mediation has been gaining in popularity since the early 2000s as an alternative method for assessing mediation [1]

What is (Causal) Mediation Analysis?

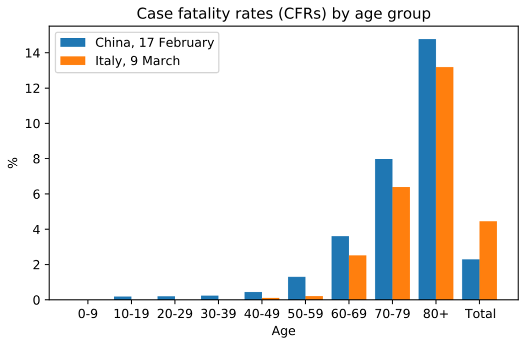
- Mediation analysis is the study of how a treatment X influences an outcome Y **through one or more intermediate variables**, called mediators M
- Mediation analysis enables us to understand **the pathways by which a cause produces an effect**
- Causal mediation analysis performs such analysis through the lenses of **Causal Inference framework**
- While traditional mediation is prevalent in the literature, causal mediation has been gaining in popularity since the early 2000s as an alternative method for assessing mediation [1]

What is (Causal) Mediation Analysis?

- Mediation analysis is the study of how a treatment X influences an outcome Y **through one or more intermediate variables**, called mediators M
- Mediation analysis enables us to understand **the pathways by which a cause produces an effect**
- Causal mediation analysis performs such analysis through the lenses of **Causal Inference framework**
- While traditional mediation is prevalent in the literature, causal mediation has been gaining in popularity since the early 2000s as an alternative method for assessing mediation [1]

Our Running Example

Case Fatality Rate (CFR): proportion of confirmed cases that end fatally

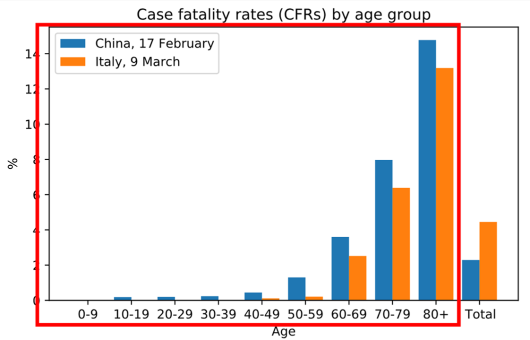


- Problem:
  - **for all age groups** CFRs in Italy are **lower** than those in China,
  - but **the total** CFR in Italy is **higher** than that in China
- Which country was "better" at managing the pandemic?
- Simpson's Paradox[2]: data tells different stories when analyzed overall vs. by age group

Credits: example from von Kügelgen et al., "Simpson's paradox in COVID-19 case fatality rates: A mediation analysis of age-related causal effects"[2]

Our Running Example

Case Fatality Rate (CFR): proportion of confirmed cases that end fatally

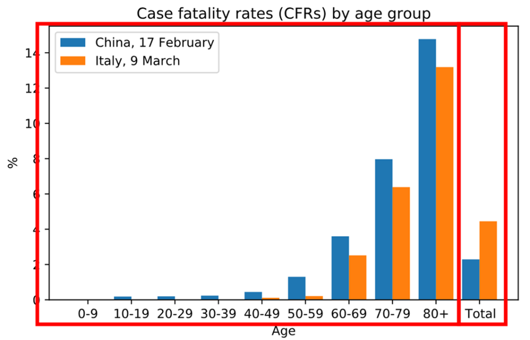


- Problem:
  - **for all age groups** CFRs in Italy are **lower** than those in China,
  - but **the total** CFR in Italy is **higher** than that in China
- Which country was "better" at managing the pandemic?
- Simpson's Paradox[2]: data tells different stories when analyzed overall vs. by age group

Credits: example from von Kügelgen et al., "Simpson's paradox in COVID-19 case fatality rates: A mediation analysis of age-related causal effects"[2]

Our Running Example

Case Fatality Rate (CFR): proportion of confirmed cases that end fatally

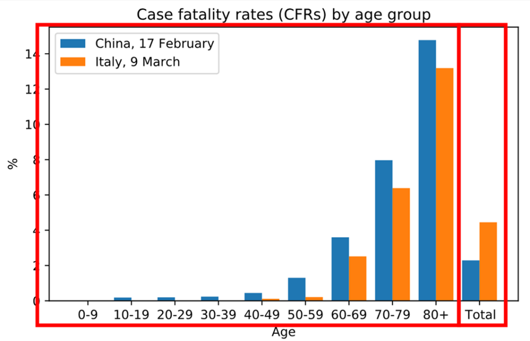


- Problem:
  - **for all age groups** CFRs in Italy are **lower** than those in China,
  - but **the total** CFR in Italy is **higher** than that in China
- Which country was “better” at managing the pandemic?
- Simpson’s Paradox[2]: data tells different stories when analyzed overall vs. by age group

Credits: example from von Kügelgen et al., "Simpson’s paradox in COVID-19 case fatality rates: A mediation analysis of age-related causal effects"[2]

Our Running Example

Case Fatality Rate (CFR): proportion of confirmed cases that end fatally

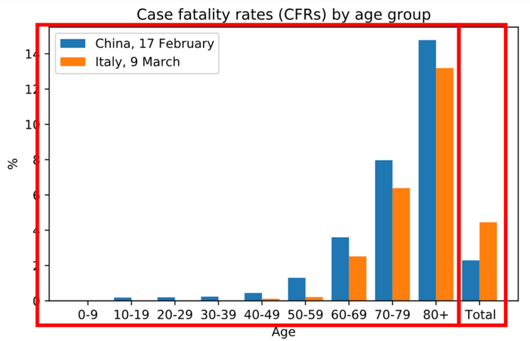


- Problem:
  - **for all age groups** CFRs in Italy are **lower** than those in China,
  - but **the total** CFR in Italy is **higher** than that in China
- Which country was "better" at managing the pandemic?
- Simpson's Paradox[2]: data tells different stories when analyzed overall vs. by age group

Credits: example from von Kügelgen et al., "Simpson's paradox in COVID-19 case fatality rates: A mediation analysis of age-related causal effects"[2]

Our Running Example

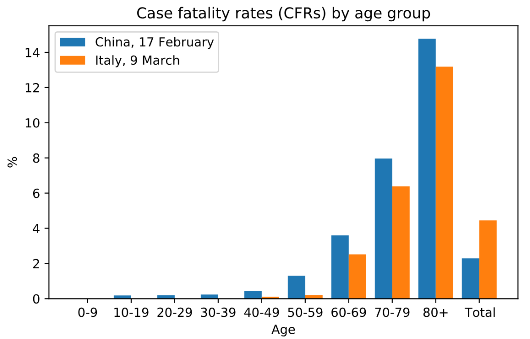
Case Fatality Rate (CFR): proportion of confirmed cases that end fatally



- Problem:
  - **for all age groups** CFRs in Italy are **lower** than those in China,
  - but **the total** CFR in Italy is **higher** than that in China
- Which country was "better" at managing the pandemic?
- Simpson's Paradox[2]: data tells different stories when analyzed overall vs. by age group

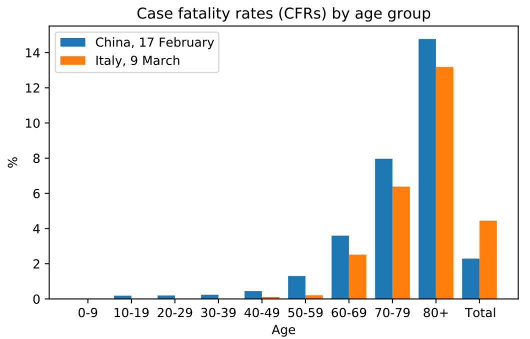
Credits: example from von Kügelgen et al., "Simpson's paradox in COVID-19 case fatality rates: A mediation analysis of age-related causal effects"[2]

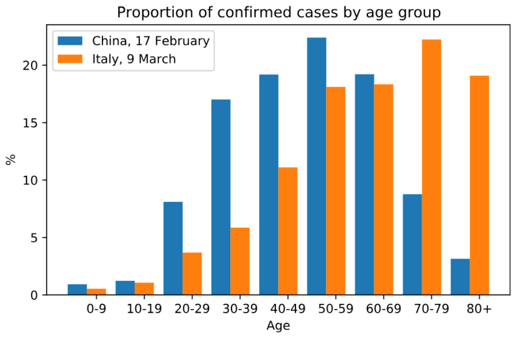
Our Running Example



How can such a pattern be explained?
CFRs are relative frequencies!

Our Running Example



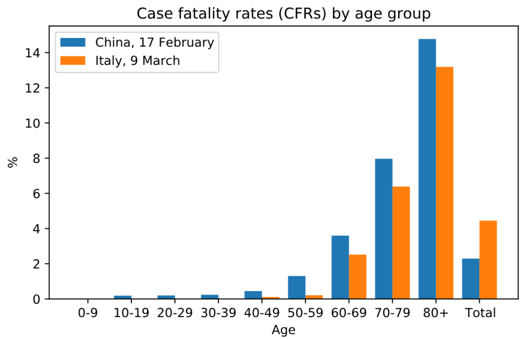


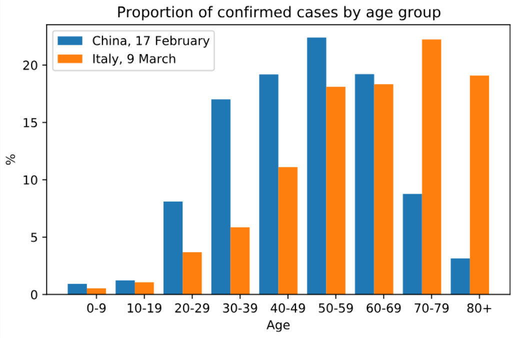
How can such a pattern be explained?
CFRs are relative frequencies!

We must relate it to absolute number of cases:

- Case age distribution differed by country
- Italy: majority aged 60+ (higher mortality risk)
- China: majority aged 30–59

Our Running Example





How can such a pattern be explained?
CFRs are relative frequencies!

We must relate it to absolute number of cases:

- Case age distribution differed by country
- Italy: majority aged 60+ (higher mortality risk)
- China: majority aged 30–59

Our Running Example

Conclusion

1. The Italian population is older than the Chinese one
2. Italy had a larger share of confirmed cases among elderly people
3. The elderly are generally at higher risk when contracting COVID-19 effect
4. This explains the Simpson's paradox in the data

We will now see how this example can be studied through the lenses of causal mediation analysis for a transparent comparison of CFRs across countries

Our Running Example

Conclusion

1. The Italian population is older than the Chinese one
2. Italy had a larger share of confirmed cases among elderly people
3. The elderly are generally at higher risk when contracting COVID-19 effect
4. This explains the Simpson's paradox in the data

We will now see how this example can be studied through the lenses of causal mediation analysis for a transparent comparison of CFRs across countries

Our Running Example

Conclusion

1. The Italian population is older than the Chinese one
2. Italy had a larger share of confirmed cases among elderly people
3. The elderly are generally at higher risk when contracting COVID-19 effect
4. This explains the Simpson's paradox in the data

We will now see how this example can be studied through the lenses of causal mediation analysis for a transparent comparison of CFRs across countries

Our Running Example

Conclusion

1. The Italian population is older than the Chinese one
2. Italy had a larger share of confirmed cases among elderly people
3. The elderly are generally at higher risk when contracting COVID-19 effect
4. This explains the Simpson's paradox in the data

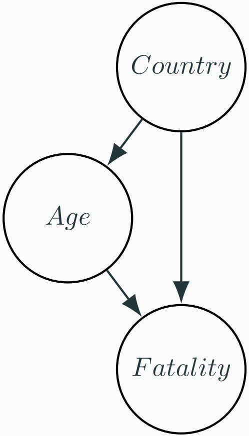
We will now see how this example can be studied through the lenses of causal mediation analysis for a transparent comparison of CFRs across countries

Conclusion

1. The Italian population is older than the Chinese one
2. Italy had a larger share of confirmed cases among elderly people
3. The elderly are generally at higher risk when contracting COVID-19 effect
4. This explains the Simpson's paradox in the data

We will now see how this example can be studied through the lenses of causal mediation analysis for a transparent comparison of CFRs across countries

Causal Graph

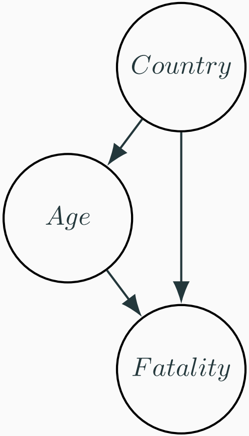


This is a simple and coarse-grained view of a complex underlying phenomenon:

- **(Country → Age)** the case demographic is country-dependent
- **(Age → Fatality)** COVID-19 is more dangerous for the elderly
- **(Country → Fatality)** summarizes country-specific influences on case fatality **other than age** (medical infrastructure, availability of hospital beds and ventilators, local expertise, pandemic-preparedness, air pollution levels, etc.)

Causal Mediation Analysis

Causal Graph

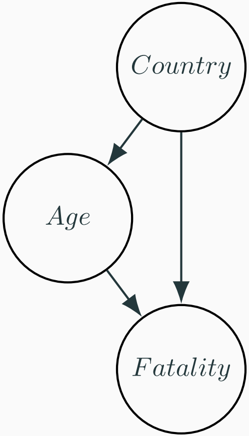


This is a simple and coarse-grained view of a complex underlying phenomenon:

- **(Country → Age)** the case demographic is country-dependent
- **(Age → Fatality)** COVID-19 is more dangerous for the elderly
- **(Country → Fatality)** summarizes country-specific influences on case fatality **other than age** (medical infrastructure, availability of hospital beds and ventilators, local expertise, pandemic-preparedness, air pollution levels, etc.)

Causal Mediation Analysis

Causal Graph

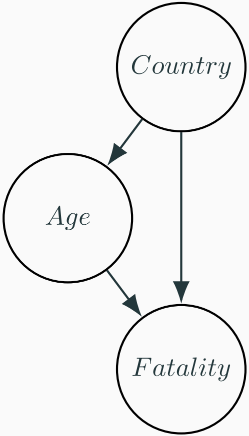


This is a simple and coarse-grained view of a complex underlying phenomenon:

- **(Country → Age)** the case demographic is country-dependent
- **(Age → Fatality)** COVID-19 is more dangerous for the elderly
- **(Country → Fatality)** summarizes country-specific influences on case fatality **other than age** (medical infrastructure, availability of hospital beds and ventilators, local expertise, pandemic-preparedness, air pollution levels, etc.)

Causal Mediation Analysis

Causal Graph



This is a simple and coarse-grained view of a complex underlying phenomenon:

- **(Country → Age)** the case demographic is country-dependent
- **(Age → Fatality)** COVID-19 is more dangerous for the elderly
- **(Country → Fatality)** summarizes country-specific influences on case fatality **other than age** (medical infrastructure, availability of hospital beds and ventilators, local expertise, pandemic-preparedness, air pollution levels, etc.)

A couple of notes

- Stay focused on the semantics/intuition so forget about assumptions like causal sufficiency, positivity, etc. (even if they are important!)

A couple of notes

- Stay focused on the semantics/intuition so forget about assumptions like causal sufficiency, positivity, etc. (even if they are important!)
- Even though we are talking about causality, we are not obliged to stick with a "mechanistic view"

A couple of notes

- Stay focused on the semantics/intuition so forget about assumptions like causal sufficiency, positivity, etc. (even if they are important!)
- Even though we are talking about causality, we are not obliged to stick with a "mechanistic view"

Indeed in our example we are

A couple of notes

- Stay focused on the semantics/intuition so forget about assumptions like causal sufficiency, positivity, etc. (even if they are important!)
- Even though we are talking about causality, we are not obliged to stick with a "mechanistic view"

Indeed in our example we are

- not concerned with how controlling X may bring about change in Y,

A couple of notes

- Stay focused on the semantics/intuition so forget about assumptions like causal sufficiency, positivity, etc. (even if they are important!)
- Even though we are talking about causality, we are not obliged to stick with a "mechanistic view"

Indeed in our example we are

- not concerned with how controlling X may bring about change in Y,
- but in understanding how natural variations of X affect the outcome Y

A couple of notes

- Stay focused on the semantics/intuition so forget about assumptions like causal sufficiency, positivity, etc. (even if they are important!)
- Even though we are talking about causality, we are not obliged to stick with a "mechanistic view"

Indeed in our example we are

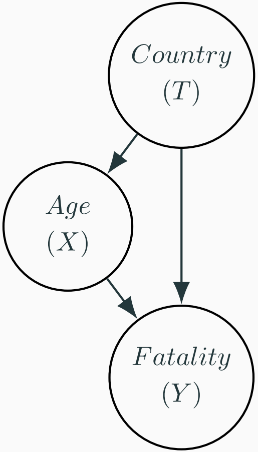
- not concerned with how controlling X may bring about change in Y,
- but in understanding how natural variations of X affect the outcome Y
- we cannot force a change from country X to country Y!

Causal Mediation Queries

We can compute four types of effects:

- **Total Causal Effect**
- **Controlled Direct Effect**
- **Natural Direct Effect**
- **Natural Indirect Effect**

We will see their causal estimands, Pearl provides theorems for identification in [3]



Total Causal Effect (TCE)

Question:

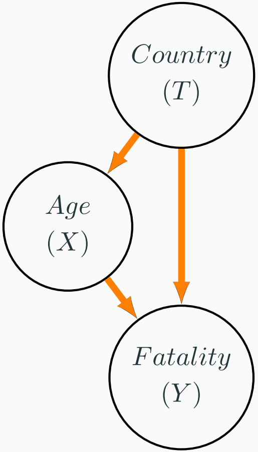
"What would be the (overall) effect on fatality of changing country from China to Italy?"

Causal estimand:

$$\begin{aligned} \mathrm{TCE}_{0\to 1} = &\mathbb{E}[Y|do(T=1)] \\ &- \mathbb{E}[Y|do(T=0)] \end{aligned}$$

0:China; 1:Italy

Causal Mediation Analysis



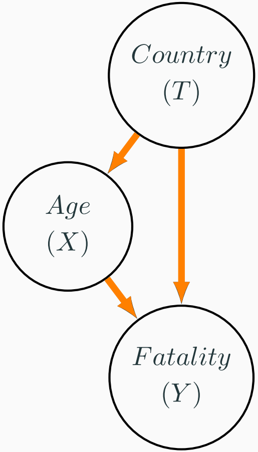
Total Causal Effect (TCE)

Question:
"What would be the (overall) effect on fatality of changing country from China to Italy?"

Causal estimand:

$$\begin{aligned}\mathrm{TCE}_{0\to1} =& \mathbb{E}[Y|do(T=1)] \\ &- \mathbb{E}[Y|do(T=0)]\end{aligned}$$

0:China; 1:Italy



Total Causal Effect (TCE)

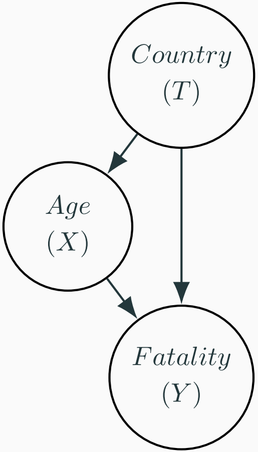
Question:

"What would be the (overall) effect on fatality of changing country from China to Italy?"

Causal estimand:

$$\begin{aligned}\mathrm{TCE}_{0\to 1} =&\mathbb{E}[Y|do(T=1)]\\ &-\mathbb{E}[Y|do(T=0)]\end{aligned}$$

0:China; 1:Italy



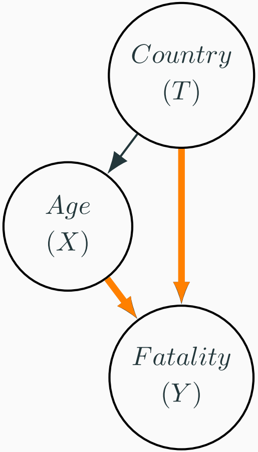
Controlled Direct Effect (CDE)

Question:

"For 50–59 year olds, in which country is it safer to get the disease?"

Causal estimand:

$$\begin{aligned}\mathrm{CDE}_{0\to 1} = & \mathbb{E}[Y|do(T=1, X=x)] \\ & - \mathbb{E}[Y|do(T=0, X=x)]\end{aligned}$$

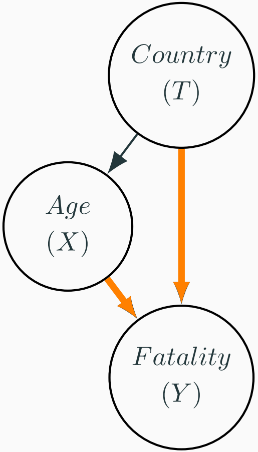


Controlled Direct Effect (CDE)

Question:
"For 50–59 year olds, in which country is it safer to get the disease?"

Causal estimand:

$$\begin{aligned}\mathrm{CDE}_{0\to 1} =& \mathbb{E}[Y|do(T=1, X=x)] \\ & - \mathbb{E}[Y|do(T=0, X=x)]\end{aligned}$$



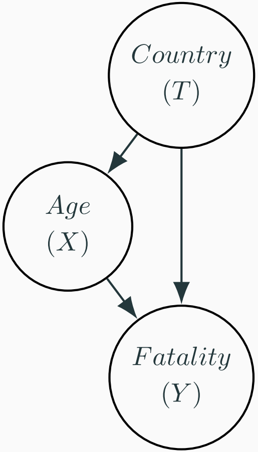
Controlled Direct Effect (CDE)

Question:
"For 50–59 year olds, in which country is it safer to get the disease?"

Causal estimand:

$$\begin{aligned}\mathrm{CDE}_{0\to 1} =& \mathbb{E}[Y|do(T=1, X=x)] \\ &- \mathbb{E}[Y|do(T=0, X=x)]\end{aligned}$$

Causal Mediation Analysis



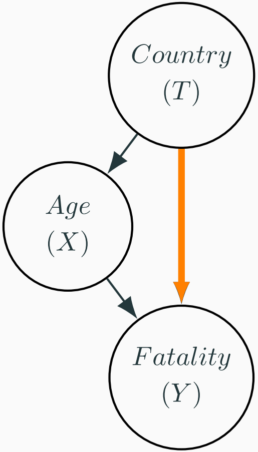
Natural Direct Effect (NDE)

Question:
"For the Chinese case demographic, would the Italian approach have been better?"

Causal estimand:

$$\begin{aligned}\mathrm{NDE}_{0\to 1} = &\mathbb{E}[Y_{X(0)}|do(T=1)] \\ &- \mathbb{E}[Y|do(T=0)]\end{aligned}$$

X(0) is the counterfactual of X had T been 0



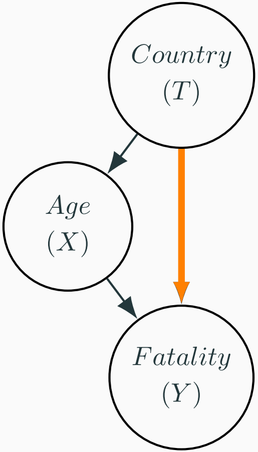
Natural Direct Effect (NDE)

Question:
"For the Chinese case demographic, would the Italian approach have been better?"

Causal estimand:

$$\begin{aligned} \mathrm{NDE}_{0\to 1} = & \mathbb{E}[Y_{X(0)} | do(T=1)] \\ & - \mathbb{E}[Y | do(T=0)] \end{aligned}$$

X(0) is the counterfactual of X had T been 0



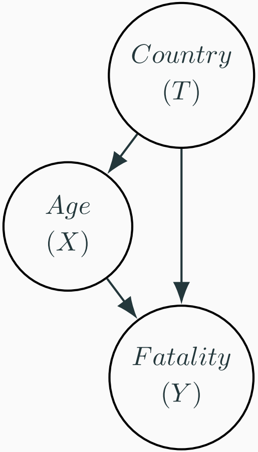
Natural Direct Effect (NDE)

Question:
"For the Chinese case demographic, would the Italian approach have been better?"

Causal estimand:

$$\begin{aligned}\mathrm{NDE}_{0\to 1} =& \mathbb{E}[Y_{X(0)}|do(T=1)] \\ & - \mathbb{E}[Y|do(T=0)]\end{aligned}$$

X(0) is the counterfactual of X had T been 0



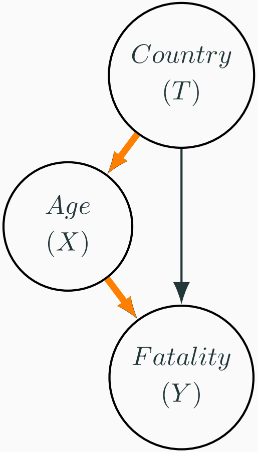
Natural Indirect Effect (NIE)

Question:

"How would the overall CFR in China change if the case demographic had instead been that from Italy while keeping all else (i.e., CFR's of each age group) the same?"

Causal estimand:

$$\begin{aligned}\mathrm{NIE}_{0\to 1} = &\mathbb{E}[Y_{X(1)} | do(T=0)] \\ &- \mathbb{E}[Y | do(T=0)]\end{aligned}$$

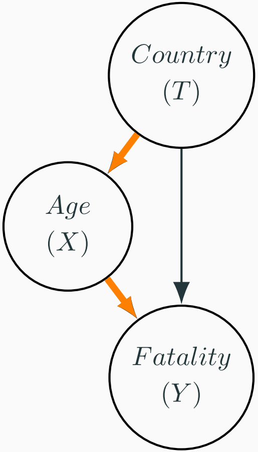


Natural Indirect Effect (NIE)

Question:
"How would the overall CFR in China change if the case demographic had instead been that from Italy while keeping all else (i.e., CFR's of each age group) the same?"

Causal estimand:

$$\mathrm{NIE}_{0\to 1} = \mathbb{E}[Y_{X(1)} | do(T=0)] - \mathbb{E}[Y | do(T=0)]$$



Natural Indirect Effect (NIE)

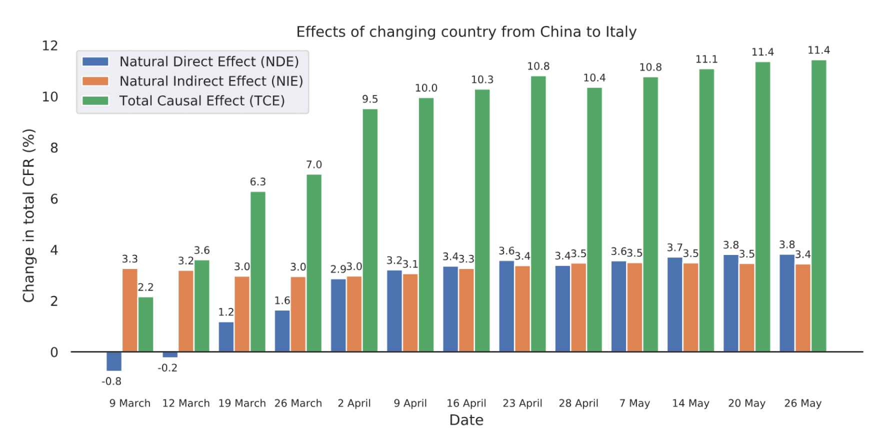
Question:

"How would the overall CFR in China change if the case demographic had instead been that from Italy while keeping all else (i.e., CFR's of each age group) the same?"

Causal estimand:

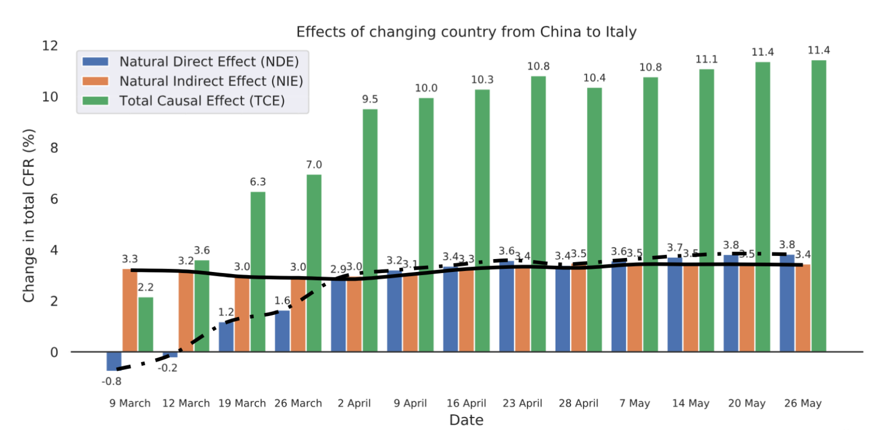
$$\begin{aligned}\text{NIE}_{0\to1} =& \mathbb{E}[Y_{X(1)}|do(T=0)] \\ &- \mathbb{E}[Y|do(T=0)]\end{aligned}$$

Simpson's Paradox through Causal Mediation Queries



- The Simpson's paradox is reflected in the **opposite signs of NDE and NIE** at the beginning
- Over time, it is mainly the NDE that drives the observed changes over time

Simpson's Paradox through Causal Mediation Queries



- The Simpson's paradox is reflected in the **opposite signs of NDE and NIE** at the beginning
- Over time, it is mainly the NDE that drives the observed changes over time

An application of Causal Mediation Analysis for Fairness Assessment

Fairness Assessment

Under algorithmic fairness [4]:

- we aim to uncover discriminatory biases of models
- we frame **discrimination** as a causal influence of a protected attribute X (such as age, sex, ethnicity, etc.) on an outcome Y of interest along **paths that are considered unfair** in the specific context
- (again) we study how natural variations of X affect the outcome Y

Fairness Assessment

Under algorithmic fairness [4]:

- we aim to uncover discriminatory biases of models
- we frame **discrimination** as a causal influence of a protected attribute X (such as age, sex, ethnicity, etc.) on an outcome Y of interest along **paths that are considered unfair** in the specific context
- (again) we study how natural variations of X affect the outcome Y

Fairness Assessment

Under algorithmic fairness [4]:

- we aim to uncover discriminatory biases of models
- we frame **discrimination** as a causal influence of a protected attribute X (such as age, sex, ethnicity, etc.) on an outcome Y of interest along **paths that are considered unfair** in the specific context
- (again) we study how natural variations of X affect the outcome Y

Triage Discrimination: Myth or Reality?

Justin Armanini[1,4,*] Fabio Stella[1,2,3]

[1]University of Milano-Bicocca, Viale Sarca 336, 20126, Milan, Italy

[2]BReCHS – Bicocca Research Centre in Health Services, Edificio U7 - Piazza dell'Ateneo Nuovo 1, 20126, Milan, Italy

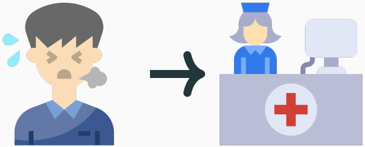
[3]Bicocca Bioinformatics Biostatistics and Bioimaging Centre – B4, Via Follereau 3, 20854, Vedano al Lambro (MB), Italy

[4]Department of Epidemiology and Data Science, Fondazione IRCCS Istituto Nazionale dei Tumori, Via Giacomo Venezian 1, 20133, Milan, Italy

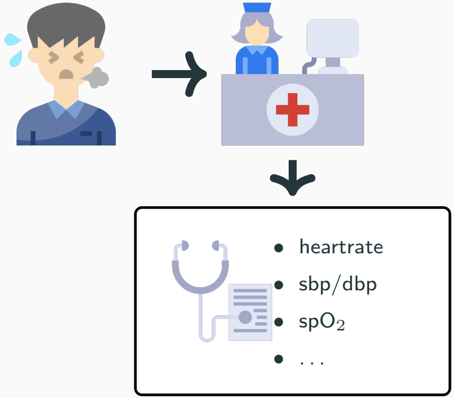
[*]Corresponding author: j.armanini@campus.unimib.it

What is Clinical Triage?

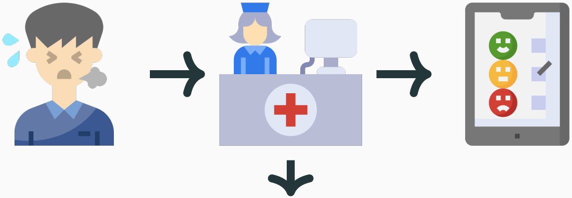
What is Clinical Triage?

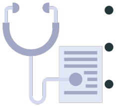


What is Clinical Triage?



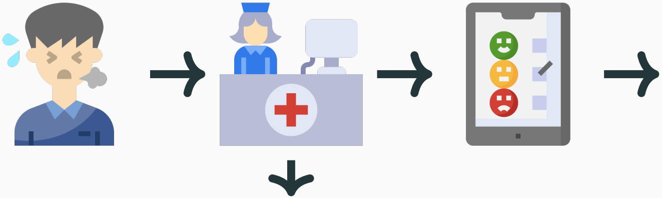
What is Clinical Triage?





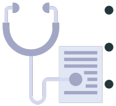
- heartrate
- sbp/dbp
- spO_2
- ...

What is Clinical Triage?



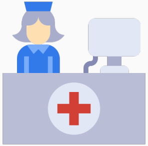
Emergency Severity Index (ESI)

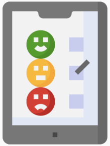
- ESI 1 (Immediate)
- ESI 2 (Emergent)
- ESI 3 (Urgent)
- ESI 4 (Nonurgent)
- ESI 5 (Minor)

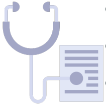


- heartrate
- sbp/dbp
- spO_2
- ...

What is Clinical Triage?







- heartrate
- sbp/dbp
- spO_2
- ...

Emergency Severity Index (ESI)

- ESI 1 (Immediate)
- ESI 2 (Emergent)
- ESI 3 (Urgent)
- ESI 4 (Nonurgent)
- ESI 5 (Minor)

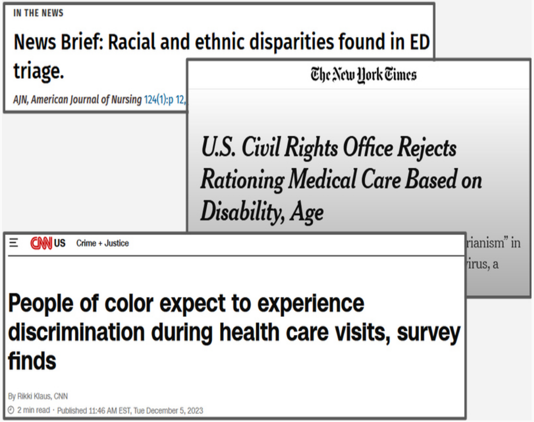
Lower ESI $\rightarrow$ Higher acuity

Problem: Allegations of Discrimination

- Documented statistical **disparities** in care according to race, gender, disability of patients
- Such disparities can lead to allegations of **discrimination**
- **Are these statistical demographic disparities evidence of actual discrimination?**

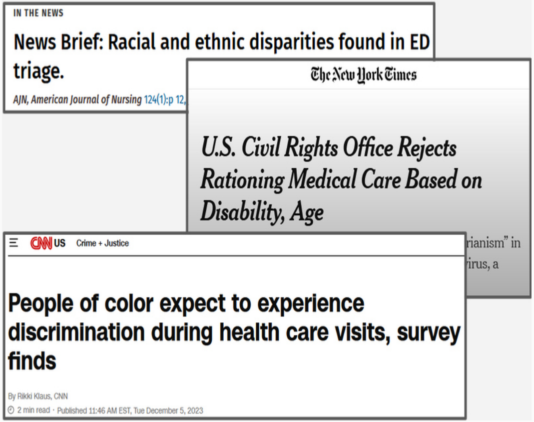
Problem: Allegations of Discrimination

- Documented statistical **disparities** in care according to race, gender, disability of patients
- Such disparities can lead to allegations of **discrimination**
- Are these statistical demographic disparities evidence of actual discrimination?



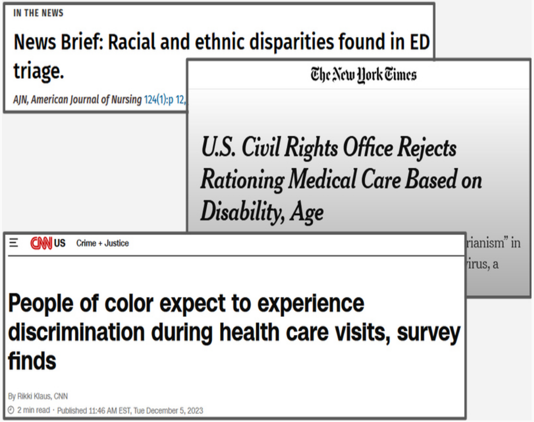
Problem: Allegations of Discrimination

- Documented statistical **disparities** in care according to race, gender, disability of patients
- Such disparities can lead to allegations of **discrimination**
- **Are these statistical demographic disparities evidence of actual discrimination?**



Problem: Allegations of Discrimination

- Documented statistical **disparities** in care according to race, gender, disability of patients
- Such disparities can lead to allegations of **discrimination**
- **Are these statistical demographic disparities evidence of actual discrimination?**


IN THE NEWS

News Brief: Racial and ethnic disparities found in ED triage.

AJN, American Journal of Nursing 124(1):p 12,

The New York Times

U.S. Civil Rights Office Rejects Rationing Medical Care Based on Disability, Age

rianism" in
irus, a

CNN US Crime + Justice

People of color expect to experience discrimination during health care visits, survey finds

By Rikki Klaus, CNN
2 min read · Published 11:46 AM EST, Tue December 5, 2023

We use the term 'race' as it is used in the relevant literature—as a socially constructed category—and only in a neutral, descriptive sense!

Why is Discrimination a Concern?

- Delayed diagnosis
- Inadequate pain management
- Increased mortality rates
- Decreased trust in medical institutions
- Violations of fundamental principles of medical ethics and social justice

Why is Discrimination a Concern?

- Delayed diagnosis
- Inadequate pain management
- Increased mortality rates
- Decreased trust in medical institutions
- Violations of fundamental principles of medical ethics and social justice

Why is Discrimination a Concern?

- Delayed diagnosis
- Inadequate pain management
- Increased mortality rates
- Decreased trust in medical institutions
- Violations of fundamental principles of medical ethics and social justice

Why is Discrimination a Concern?

- Delayed diagnosis
- Inadequate pain management
- Increased mortality rates
- Decreased trust in medical institutions
- Violations of fundamental principles of medical ethics and social justice

Why is Discrimination a Concern?

- Delayed diagnosis
- Inadequate pain management
- Increased mortality rates
- Decreased trust in medical institutions
- Violations of fundamental principles of medical ethics and social justice

Research Questions

RQ1: Does discrimination exists?

- Is there disparity in triage based on demographics such as race, gender, and age?

RQ1: Does discrimination exists?

- Is there disparity in triage based on demographics such as race, gender, and age?

Research Questions

RQ1: Does discrimination exists?

- Is there disparity in triage based on demographics such as race, gender, and age?

RQ2: What drives discrimination?

- If disparity exists, what underlying mechanisms drive such disparity?
- Is there any **justifiable** discrimination based on clinical factors?

Research Questions

RQ1: Does discrimination exists?

- Is there disparity in triage based on demographics such as race, gender, and age?

RQ2: What drives discrimination?

- If disparity exists, what underlying mechanisms drive such disparity?
- Is there any **justifiable** discrimination based on clinical factors?

Research Questions

RQ1: Does discrimination exists?

- Is there disparity in triage based on demographics such as race, gender, and age?

RQ2: What drives discrimination?

- If disparity exists, what underlying mechanisms drive such disparity?
- Is there any **justifiable** discrimination based on clinical factors?

Related Work

- Documented racial/gender disparities **in the US**
- Used **traditional statistical methods**
    - Odds Ratios (ORs) with adjustment or matching strategies
- Cannot distinguish correlation from causation
    - **Meaning**: statistical disparity doesn't prove discrimination
    - **Confounding**: adjusting is not based on the causal structure of the problem
    - **Mediation** pathways: need to disentangle direct vs. indirect causation

Related Work

- Documented racial/gender disparities **in the US**
- Used **traditional statistical methods**
  - Odds Ratios (ORs) with adjustment or matching strategies
- Cannot distinguish correlation from causation
  - **Meaning**: statistical disparity doesn't prove discrimination
  - **Confounding**: adjusting is not based on the causal structure of the problem
  - **Mediation** pathways: need to disentangle direct vs. indirect causation

Related Work

- Documented racial/gender disparities **in the US**
- Used **traditional statistical methods**
    - Odds Ratios (ORs) with adjustment or matching strategies
- Cannot distinguish correlation from causation
    - **Meaning**: statistical disparity doesn't prove discrimination
    - **Confounding**: adjusting is not based on the causal structure of the problem
    - **Mediation** pathways: need to disentangle direct vs. indirect causation

Related Work

- Documented racial/gender disparities **in the US**
- Used **traditional statistical methods**
  - Odds Ratios (ORs) with adjustment or matching strategies
- Cannot distinguish correlation from causation
  - **Meaning**: statistical disparity doesn't prove discrimination
  - **Confounding**: adjusting is not based on the causal structure of the problem
  - **Mediation** pathways: need to disentangle direct vs. indirect causation

Related Work

- Documented racial/gender disparities **in the US**
- Used **traditional statistical methods**
  - Odds Ratios (ORs) with adjustment or matching strategies
- Cannot distinguish correlation from causation
  - **Meaning**: statistical disparity doesn't prove discrimination
  - **Confounding**: adjusting is not based on the causal structure of the problem
  - **Mediation** pathways: need to disentangle direct vs. indirect causation

Related Work

- Documented racial/gender disparities **in the US**
- Used **traditional statistical methods**
  - Odds Ratios (ORs) with adjustment or matching strategies
- Cannot distinguish correlation from causation
  - **Meaning**: statistical disparity doesn't prove discrimination
  - **Confounding**: adjusting is not based on the causal structure of the problem
  - **Mediation** pathways: need to disentangle direct vs. indirect causation

Related Work

- Documented racial/gender disparities **in the US**
- Used **traditional statistical methods**
  - Odds Ratios (ORs) with adjustment or matching strategies
- Cannot distinguish correlation from causation
  - **Meaning**: statistical disparity doesn't prove discrimination
  - **Confounding**: adjusting is not based on the causal structure of the problem
  - **Mediation** pathways: need to disentangle direct vs. indirect causation

Related Work

- Documented racial/gender disparities **in the US**
- Used **traditional statistical methods**
  - Odds Ratios (ORs) with adjustment or matching strategies
- Cannot distinguish correlation from causation
  - **Meaning**: statistical disparity doesn't prove discrimination
  - **Confounding**: adjusting is not based on the causal structure of the problem
  - **Mediation** pathways: need to disentangle direct vs. indirect causation

Our original contribution: first application of causal mediation analysis to triage discrimination

Dataset

MIMIC-IV-ED (v2.2)

- **Source**: Beth Israel Deaconess Medical Center, Boston (US)
- **Time period**: 2011-2019
- **Sample size**: 160k patients

Variables

- **Demographics**: Age, Gender, Race
- **Clinical**: temperature, heart rate, respiratory rate, spO2, sbp, dbp, pain level
- **Outcome**: ESI Acuity Score (1-5)

Preprocessing

- **Discretized continuous variables** according to the healthcare literature
- Extracted first **Emergency Department** visit per patient

Dataset

MIMIC-IV-ED (v2.2)

- **Source**: Beth Israel Deaconess Medical Center, Boston (US)
- **Time period**: 2011-2019
- **Sample size**: 160k patients

Variables

- **Demographics**: Age, Gender, Race
- **Clinical**: temperature, heart rate, respiratory rate, spO2, sbp, dbp, pain level
- **Outcome**: ESI Acuity Score (1-5)

Preprocessing

- **Discretized continuous variables** according to the healthcare literature
- Extracted first **Emergency Department** visit per patient

Dataset

MIMIC-IV-ED (v2.2)

- **Source**: Beth Israel Deaconess Medical Center, Boston (US)
- **Time period**: 2011-2019
- **Sample size**: 160k patients

Variables

- **Demographics**: Age, Gender, Race
- **Clinical**: temperature, heart rate, respiratory rate, spO2, sbp, dbp, pain level
- **Outcome**: ESI Acuity Score (1-5)

Preprocessing

- **Discretized continuous variables** according to the healthcare literature
- Extracted first **Emergency Department** visit per patient

Dataset

MIMIC-IV-ED (v2.2)

- **Source**: Beth Israel Deaconess Medical Center, Boston (US)
- **Time period**: 2011-2019
- **Sample size**: 160k patients

Variables

- **Demographics**: Age, Gender, Race
- **Clinical**: temperature, heart rate, respiratory rate, spO2, sbp, dbp, pain level
- **Outcome**: ESI Acuity Score (1-5)

Preprocessing

- **Discretized continuous variables** according to the healthcare literature
- Extracted first **Emergency Department** visit per patient

Dataset

MIMIC-IV-ED (v2.2)

- **Source**: Beth Israel Deaconess Medical Center, Boston (US)
- **Time period**: 2011-2019
- **Sample size**: 160k patients

Variables

- **Demographics**: Age, Gender, Race
- **Clinical**: temperature, heart rate, respiratory rate, spO2, sbp, dbp, pain level
- **Outcome**: ESI Acuity Score (1-5)

Preprocessing

- **Discretized continuous variables** according to the healthcare literature
- Extracted first **Emergency Department** visit per patient

Dataset

MIMIC-IV-ED (v2.2)

- **Source**: Beth Israel Deaconess Medical Center, Boston (US)
- **Time period**: 2011-2019
- **Sample size**: 160k patients

Variables

- **Demographics**: Age, Gender, Race
- **Clinical**: temperature, heart rate, respiratory rate, spO2, sbp, dbp, pain level
- **Outcome**: ESI Acuity Score (1-5)

Preprocessing

- **Discretized continuous variables** according to the healthcare literature
- Extracted first **Emergency Department** visit per patient

Dataset

MIMIC-IV-ED (v2.2)

- **Source**: Beth Israel Deaconess Medical Center, Boston (US)
- **Time period**: 2011-2019
- **Sample size**: 160k patients

Variables

- **Demographics**: Age, Gender, Race
- **Clinical**: temperature, heart rate, respiratory rate, spO2, sbp, dbp, pain level
- **Outcome**: ESI Acuity Score (1-5)

Preprocessing

- **Discretized continuous variables** according to the healthcare literature
- Extracted first **Emergency Department** visit per patient

Dataset

MIMIC-IV-ED (v2.2)

- **Source**: Beth Israel Deaconess Medical Center, Boston (US)
- **Time period**: 2011-2019
- **Sample size**: 160k patients

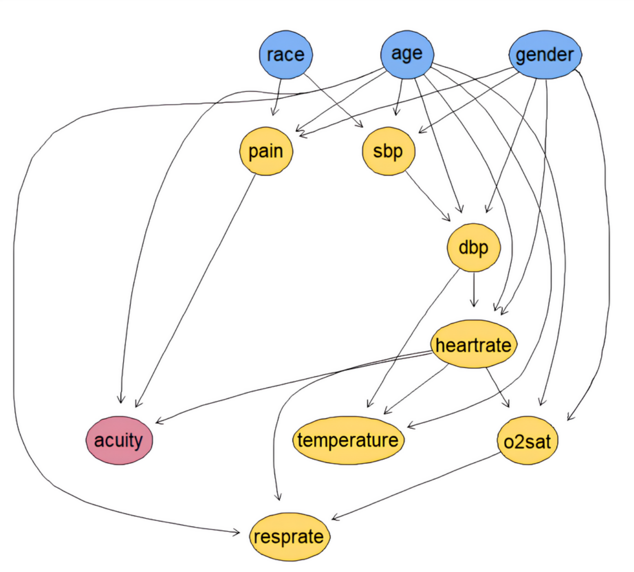
Variables

- **Demographics**: Age, Gender, Race
- **Clinical**: temperature, heart rate, respiratory rate, spO2, sbp, dbp, pain level
- **Outcome**: ESI Acuity Score (1-5)

Preprocessing

- **Discretized continuous variables** according to the healthcare literature
- Extracted first **Emergency Department** visit per patient

Learned Causal Network



Causal Network

- Directed Acyclic Graph (DAG)
- Global probability distribution over variables

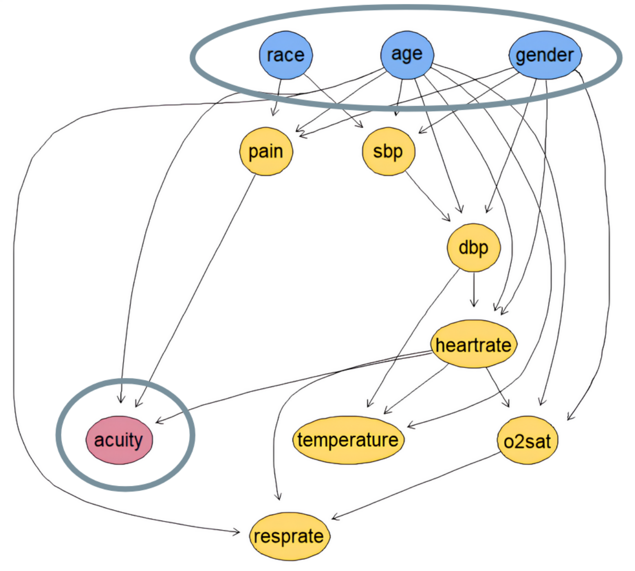
Interpretation

- Edges represent cause-effect relationships
- Distinguishes direct vs. indirect effects

Causal Learning

- Data
- Domain knowledge

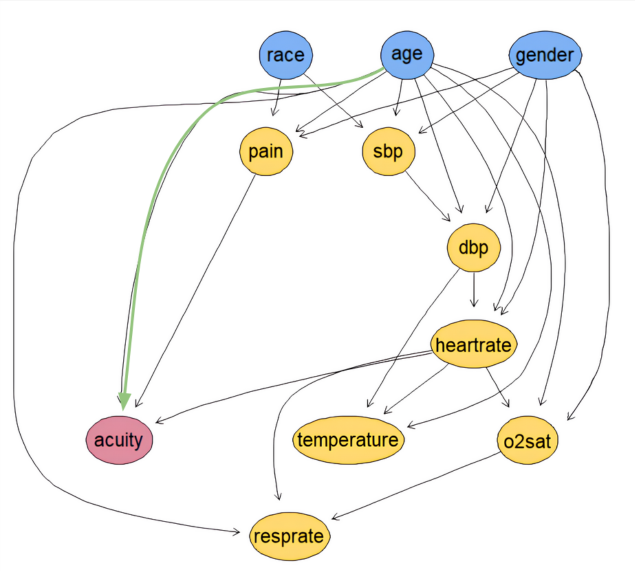
Mapping Mediation Queries



Total Causal Effect

What is the overall causal effect of a demographic variable on the triage outcome?

Mapping Mediation Queries



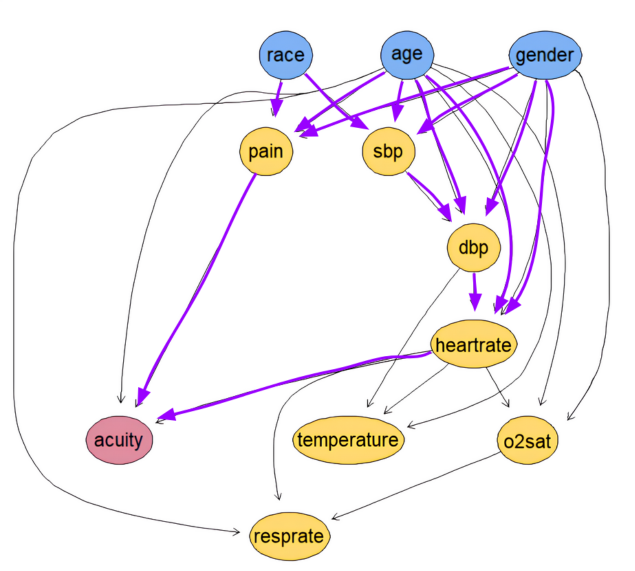
Total Causal Effect

What is the overall causal effect of a demographic variable on the triage outcome?

Natural Direct Effect

What is the direct causal effect, not mediated by clinical factors?
("**Potentially unfair disparity**" or "**Potential discrimination**")

Mapping Mediation Queries



Total Causal Effect

What is the overall causal effect of a demographic variable on the triage outcome?

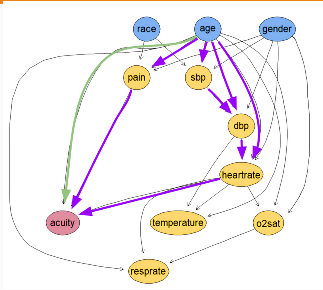
Natural Direct Effect

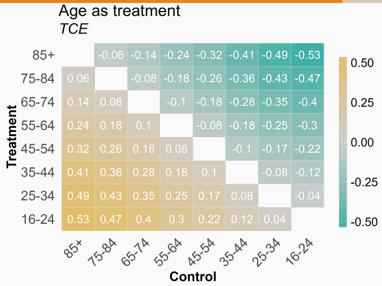
What is the direct causal effect, not mediated by clinical factors?
("**Potentially unfair disparity**" or "**Potential discrimination**")

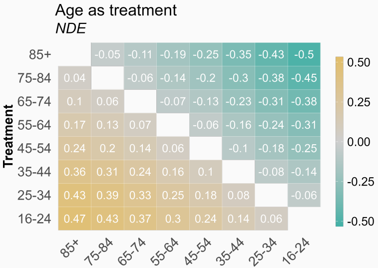
Natural Indirect Effect

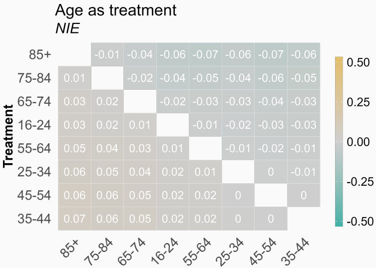
What is the causal effect mediated by clinical factors?
("**Clinically explained disparity**")

Causal Mediation Analysis: Does Discrimination Occur by Age? Why?

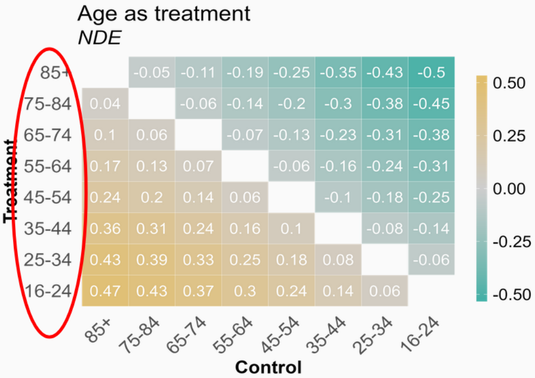




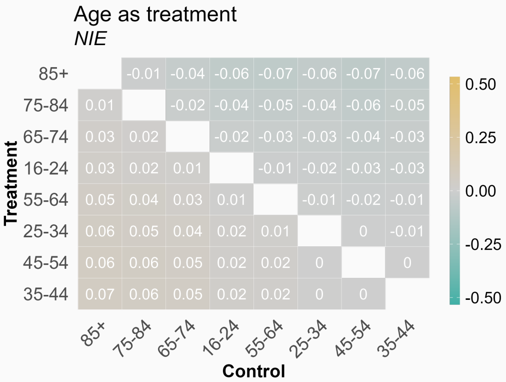




Causal Mediation Analysis: Does Discrimination Occur by Age? Why?

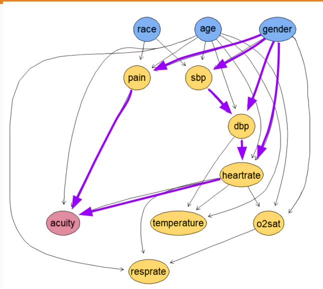


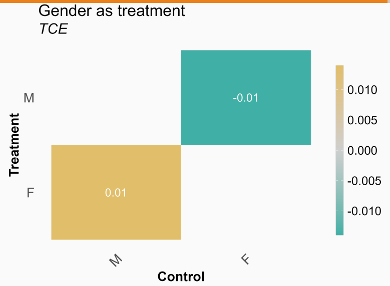
"Potentially unfair disparity"

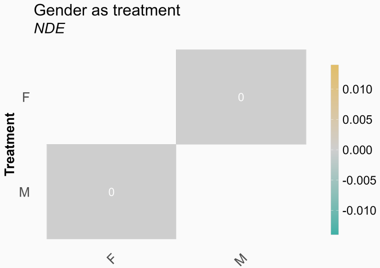


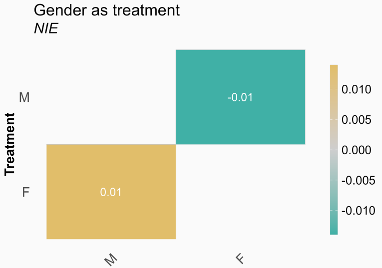
"Clinically explained disparity"

Causal Mediation Analysis: Does Discrimination Occur by Gender? Why?

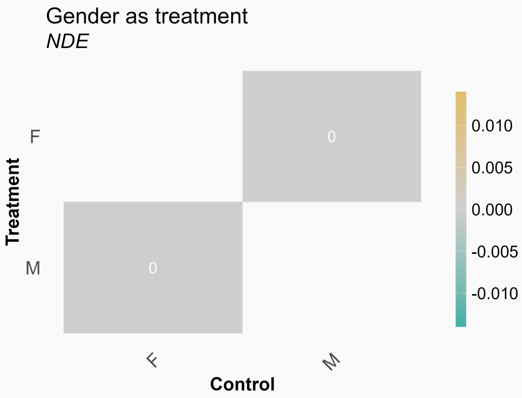




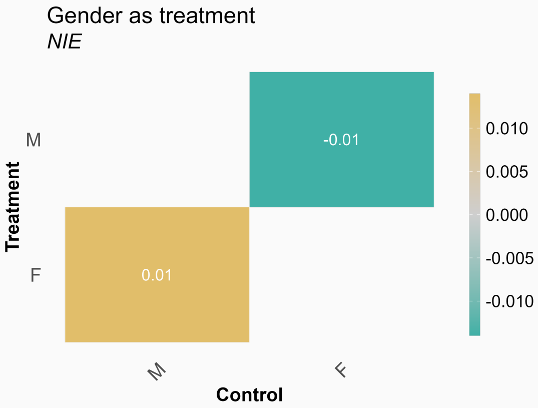




Causal Mediation Analysis: Does Discrimination Occur by Gender? Why?

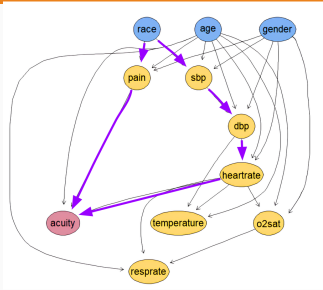


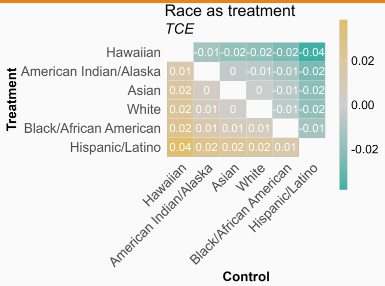
"Potentially unfair disparity"

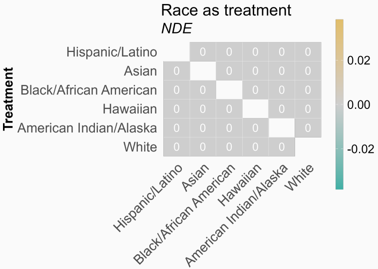


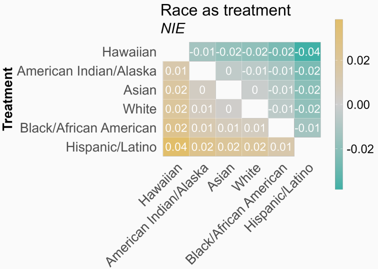
"Clinically explained disparity"

Causal Mediation Analysis: Does Discrimination Occur by Race? Why?

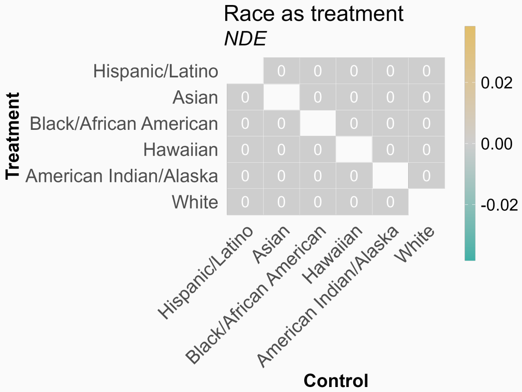


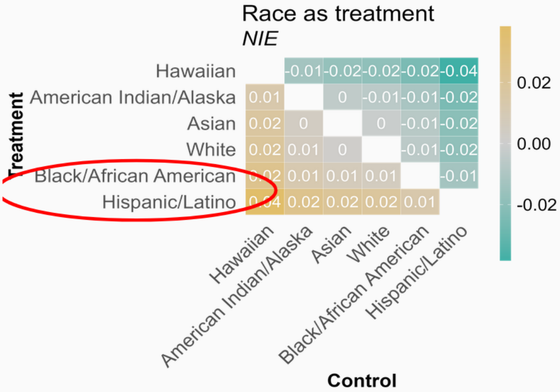






Causal Mediation Analysis: Does Discrimination Occur by Race? Why?





"Potentially unfair disparity" "Clinically explained disparity"

Conclusions

Statistical Methods

Data

Causal Mediation

Data

Conclusions

Statistical Methods

Data $\longrightarrow$ Disparity?

Causal Mediation

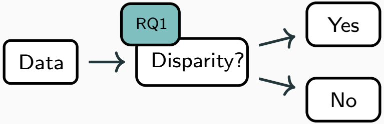
RQ1

Data $\longrightarrow$ Disparity?

Conclusions

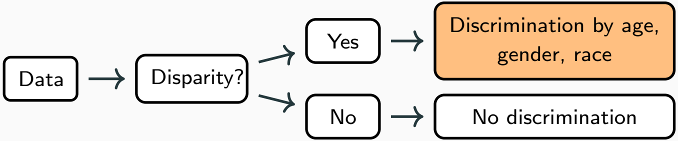
Statistical Methods

Causal Mediation

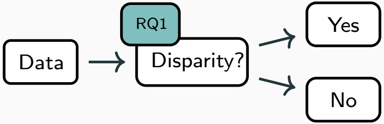


Conclusions

Statistical Methods

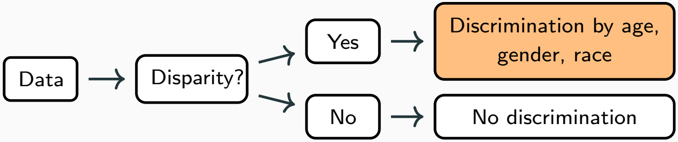


Causal Mediation

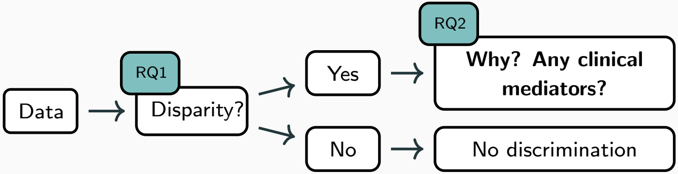


Conclusions

Statistical Methods

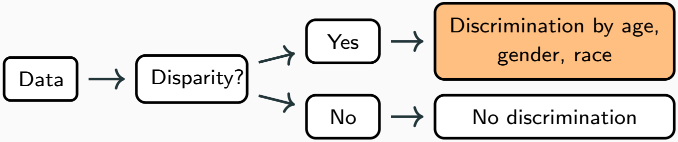


Causal Mediation

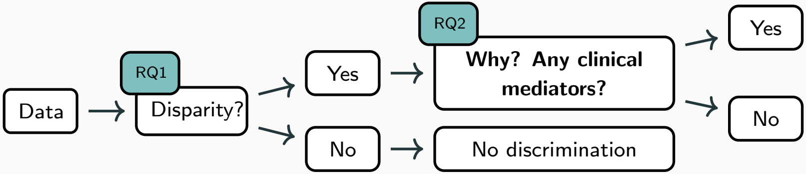


Conclusions

Statistical Methods

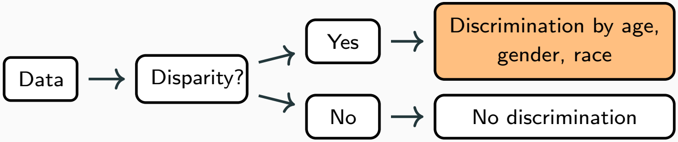


Causal Mediation

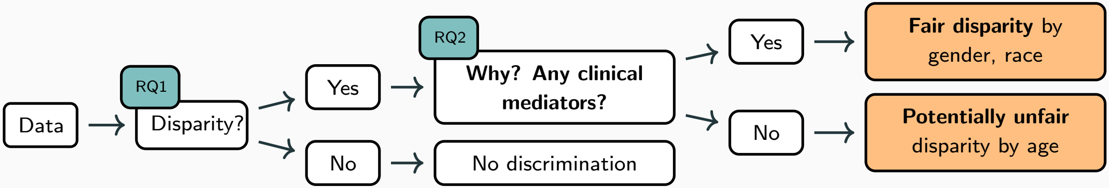


Conclusions

Statistical Methods



Causal Mediation



Key Takeaways

Theory:

- Mediation analysis decomposes total effects

Key Takeaways

Theory:

- Mediation analysis decomposes total effects
- Separates direct vs. indirect pathways

Key Takeaways

Theory:

- Mediation analysis decomposes total effects
- Separates direct vs. indirect pathways
- Causal inference framework provides rigorous modeling and identification (not discussed today, see here [3] if interested)

Key Takeaways

Theory:

- Mediation analysis decomposes total effects
- Separates direct vs. indirect pathways
- Causal inference framework provides rigorous modeling and identification (not discussed today, see here [3] if interested)

Practice:

- Applied to fairness in clinical triage

Key Takeaways

Theory:

- Mediation analysis decomposes total effects
- Separates direct vs. indirect pathways
- Causal inference framework provides rigorous modeling and identification (not discussed today, see here [3] if interested)

Practice:

- Applied to fairness in clinical triage
- Distinguish fair vs. unfair disparities

Key Takeaways

Theory:

- Mediation analysis decomposes total effects
- Separates direct vs. indirect pathways
- Causal inference framework provides rigorous modeling and identification (not discussed today, see here [3] if interested)

Practice:

- Applied to fairness in clinical triage
- Distinguish fair vs. unfair disparities
- Clinical factors explain some—but not all—disparities

Thank You for your Attention!

MADLab

The Models and Algorithms for Data & Text Mining Laboratory is a research lab at University of Milano-Bicocca focused on Causal Networks, Bayesian Networks and Continuous-Time Bayesian Networks applied to Healthcare and Medicine.

More at: https://mad.disco.unimib.it/

References i

[1] X. Qin, **"An introduction to causal mediation analysis,"** *Asia Pacific Education Review*, vol. 25, no. 3, pp. 703–717, 2024.

[2] J. von Kugelgen, L. Gresele, and B. Scholkopf, **"Simpson's paradox in COVID-19 case fatality rates: A mediation analysis of Age-Related causal effects,"** en, *IEEE Trans Artif Intell*, vol. 2, no. 1, pp. 18–27, Apr. 2021.

[3] J. Pearl, **"Interpretation and identification of causal mediation,"** en, *Psychol Methods*, vol. 19, no. 4, pp. 459–481, Jun. 2014.

[4] J. Zhang and E. Bareinboim, **"Fairness in decision-making — the causal explanation formula,"** in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'18/IAAI'18/EAAI'18, New Orleans, Louisiana, USA: AAAI Press, 2018, ISBN: 978-1-57735-800-8.