A short intro to causal inference with an example from analysis of linked registries

Jon Michael Gran

March 4, 2024

UiO : University of Oslo

What is causal inference?

- The **formal process of assessing cause and effect**

What is causal inference?

- The **formal process of assessing cause and effect**:

  "Require us to provide mathematical definitions that agree with the investigators informal and intuitive research questions"

  (loosely citing Robins 1987)

What is causal inference?

- The **formal process of assessing cause and effect**:

  "Require us to provide mathematical definitions that agree with the investigators informal and intuitive research questions"

  (loosely citing Robins 1987)

- **Not magic**

What is causal inference?

- The **formal process of assessing cause and effect**:

  "Require us to provide mathematical definitions that agree with the investigators informal and intuitive research questions"

  (loosely citing Robins 1987)
- **Not magic**: we calculate functions of observed data and interpret the output

What is causal inference?

- The **formal process of assessing cause and effect**:

  "Require us to provide mathematical definitions that agree with the investigators informal and intuitive research questions"

  (loosely citing Robins 1987)
- **Not magic**: we calculate functions of observed data and interpret the output
- **A framework** to i) ensure that the outputs answer to well defined research questions

What is causal inference?

- The **formal process of assessing cause and effect**:

  "Require us to provide mathematical definitions that agree with the investigators informal and intuitive research questions"

  (loosely citing Robins 1987)

- **Not magic**: we calculate functions of observed data and interpret the output
- **A framework** to i) ensure that the outputs answer to well defined research questions and ii) help us avoid methodological pitfalls when answering them

What is causal inference?

- The **formal process of assessing cause and effect**:

  "Require us to provide mathematical definitions that agree with the investigators informal and intuitive research questions"

  (loosely citing Robins 1987)

- **Not magic**: we calculate functions of observed data and interpret the output

- **A framework** to i) ensure that the outputs answer to well defined research questions and ii) help us avoid methodological pitfalls when answering them, iii) which have led to many new methodological developments

A classification of data analytic tasks

1. **Description** (e.g. of disease prevalence)
2. **Prediction** (e.g. for prognosis or diagnosis)
3. **Causal prediction** (e.g. for effects of medical treatments)

A classification of data analytic tasks

1. **Description** (e.g. of disease prevalence)
2. **Prediction** (e.g. for prognosis or diagnosis)
3. **Causal prediction** (e.g. for effects of medical treatments)

See e.g. Shmueli 2010 and Hernan et al. 2019

Causal versus statistical inference

Causal versus statistical inference

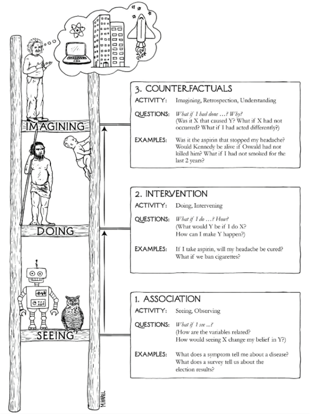
- Causal inference require **an extra layer**:

  causal assumption + statistical assumptions + data

Causal versus statistical inference

- Causal inference require **an extra layer**:

  causal assumption + statistical assumptions + data
- See also Pearl's **ladder of causation**:



Interventionism

- Causal effects are most easily defined through **interventions**

Interventionism

- Causal effects are most easily defined through **interventions**:
  What is the average outcome if everyone gets treated vs not?

Interventionism

- Causal effects are most easily defined through **interventions**:

  What is the average outcome if everyone gets treated vs not?
- Formalised through **causal estimands**:

  $$\text{ATE} = E(Y^{a=1}) - E(Y^{a=0}),$$

  for an outcome Y and treatment A taking values a

Interventionism

- Causal effects are most easily defined through **interventions**:
  What is the average outcome if everyone gets treated vs not?
- Formalised through **causal estimands**:
  $$\text{ATE} = E(Y^{a=1}) - E(Y^{a=0}),$$
  for an outcome Y and treatment A taking values a
  "No causation without manipulation" (Holland 1986)

Interventionism

- Causal effects are most easily defined through **interventions**:

  What is the average outcome if everyone gets treated vs not?

- Formalised through **causal estimands**:

  $$\text{ATE} = E(Y^{a=1}) - E(Y^{a=0}),$$

  for an outcome Y and treatment A taking values a

  "No causation without manipulation" (Holland 1986)

  Some follow **Holland's mantra** more dogmatically than others, but most would agree it simplifies things...

A history of causal inference in medicine

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**
- 1923: Jerzy Neyman introduce **counterfactuals**

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**
- 1923: Jerzy Neyman introduce **counterfactuals**
- 1925: Formal theory of **randomised trials** by Ronald Fisher

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**
- 1923: Jerzy Neyman introduce **counterfactuals**
- 1925: Formal theory of **randomised trials** by Ronald Fisher
- 1948: **First randomised clinical trial** on the effect of streptomycin for pulmonary tuberculosis, much due to statistician Bradford Hill

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**
- 1923: Jerzy Neyman introduce **counterfactuals**
- 1925: Formal theory of **randomised trials** by Ronald Fisher
- 1948: **First randomised clinical trial** on the effect of streptomycin for pulmonary tuberculosis, much due to statistician Bradford Hill
- 1954: Observational study by Doll and Bradford Hill considered to have established effect of **smoking on lung cancer**

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**
- 1923: Jerzy Neyman introduce **counterfactuals**
- 1925: Formal theory of **randomised trials** by Ronald Fisher
- 1948: **First randomised clinical trial** on the effect of streptomycin for pulmonary tuberculosis, much due to statistician Bradford Hill
- 1954: Observational study by Doll and Bradford Hill considered to have established effect of **smoking on lung cancer**
- 1974: Donald Rubin introduce counterfactuals (**potential outcomes**) in the context of observational studies

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**
- 1923: Jerzy Neyman introduce **counterfactuals**
- 1925: Formal theory of **randomised trials** by Ronald Fisher
- 1948: **First randomised clinical trial** on the effect of streptomycin for pulmonary tuberculosis, much due to statistician Bradford Hill
- 1954: Observational study by Doll and Bradford Hill considered to have established effect of **smoking on lung cancer**
- 1974: Donald Rubin introduce counterfactuals (**potential outcomes**) in the context of observational studies
- 1983: Introduction of **the propensity score** by Rosenbaum and Rubin

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**
- 1923: Jerzy Neyman introduce **counterfactuals**
- 1925: Formal theory of **randomised trials** by Ronald Fisher
- 1948: **First randomised clinical trial** on the effect of streptomycin for pulmonary tuberculosis, much due to statistician Bradford Hill
- 1954: Observational study by Doll and Bradford Hill considered to have established effect of **smoking on lung cancer**
- 1974: Donald Rubin introduce counterfactuals (**potential outcomes**) in the context of observational studies
- 1983: Introduction of **the propensity score** by Rosenbaum and Rubin
- 1986: Introduction of **g-methods** by Robins

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**
- 1923: Jerzy Neyman introduce **counterfactuals**
- 1925: Formal theory of **randomised trials** by Ronald Fisher
- 1948: **First randomised clinical trial** on the effect of streptomycin for pulmonary tuberculosis, much due to statistician Bradford Hill
- 1954: Observational study by Doll and Bradford Hill considered to have established effect of **smoking on lung cancer**
- 1974: Donald Rubin introduce counterfactuals (**potential outcomes**) in the context of observational studies
- 1983: Introduction of **the propensity score** by Rosenbaum and Rubin
- 1986: Introduction of **g-methods** by Robins
- 1995: Pearl's first full paper on **causal diagrams** (DAGs)

A history of causal inference in medicine

- 600 BC: Early description of clinical trial in **Book of Daniel**
- 1662: First known proposal of randomisation by van Helmont to compare treatment with and without **bloodletting and purging** for fever
- 1747: James Lind's **scurvy trial**
- 1923: Jerzy Neyman introduce **counterfactuals**
- 1925: Formal theory of **randomised trials** by Ronald Fisher
- 1948: **First randomised clinical trial** on the effect of streptomycin for pulmonary tuberculosis, much due to statistician Bradford Hill
- 1954: Observational study by Doll and Bradford Hill considered to have established effect of **smoking on lung cancer**
- 1974: Donald Rubin introduce counterfactuals (**potential outcomes**) in the context of observational studies
- 1983: Introduction of **the propensity score** by Rosenbaum and Rubin
- 1986: Introduction of **g-methods** by Robins
- 1995: Pearl's first full paper on **causal diagrams** (DAGs)
- 2020: **ICH E9 (R1) addendum** on estimands and sensitivity analysis in clinical trials

Observational versus randomized studies

Observational versus randomized studies

- **Randomised studies (RCTs) are the ideal** for assessing questions of causal effects

Observational versus randomized studies

- **Randomised studies (RCTs) are the ideal** for assessing questions of causal effects

  … and they should be – but they are not alway feasible

Observational versus randomized studies

- **Randomised studies (RCTs) are the ideal** for assessing questions of causal effects

  ... and they should be – but they are not alway feasible

- RCTs will often struggle with **contamination, drop-out, non-compliance, missing data and lack of generalisability**

Observational versus randomized studies

- **Randomised studies (RCTs) are the ideal** for assessing questions of causal effects

  ... and they should be – but they are not alway feasible

- RCTs will often struggle with **contamination, drop-out, non-compliance, missing data and lack of generalisability**

  These are all observational features – hence; the dichotomy between observational studies and RCTs is a false one

Observational versus randomized studies

- **Randomised studies (RCTs) are the ideal** for assessing questions of causal effects

  ... and they should be – but they are not alway feasible

- RCTs will often struggle with **contamination, drop-out, non-compliance, missing data and lack of generalisability**

  These are all observational features – hence; the dichotomy between observational studies and RCTs is a false one

- **Target trials have entered observational studies**, specifying protocols of hypothetical trial to avoid common pitfalls

Observational versus randomized studies

- **Randomised studies (RCTs) are the ideal** for assessing questions of causal effects

  ... and they should be – but they are not alway feasible

- RCTs will often struggle with **contamination, drop-out, non-compliance, missing data and lack of generalisability**

  These are all observational features – hence; the dichotomy between observational studies and RCTs is a false one

- **Target trials have entered observational studies**, specifying protocols of hypothetical trial to avoid common pitfalls

  (observational studies analysed like RCTs)

Observational versus randomized studies

- **Randomised studies (RCTs) are the ideal** for assessing questions of causal effects

  ... and they should be – but they are not alway feasible

- RCTs will often struggle with **contamination, drop-out, non-compliance, missing data and lack of generalisability**

  These are all observational features – hence; the dichotomy between observational studies and RCTs is a false one

- **Target trials have entered observational studies**, specifying protocols of hypothetical trial to avoid common pitfalls

  (observational studies analysed like RCTs)

- **Estimands have entered ICH E9 guidelines** for RCTs

Observational versus randomized studies

- **Randomised studies (RCTs) are the ideal** for assessing questions of causal effects

  ... and they should be – but they are not alway feasible

- RCTs will often struggle with **contamination, drop-out, non-compliance, missing data and lack of generalisability**

  These are all observational features – hence; the dichotomy between observational studies and RCTs is a false one

- **Target trials have entered observational studies**, specifying protocols of hypothetical trial to avoid common pitfalls

  (observational studies analysed like RCTs)

- **Estimands have entered ICH E9 guidelines** for RCTs

  (RCTs analysed like observational studies)

The causal roadmap

The causal roadmap

1. Specify **estimand** (e.g. aided by target trial protocols)

The causal roadmap

1. Specify **estimand** (e.g. aided by target trial protocols)
2. Lay out assumptions needed for **identification**
   (e.g. aided by graphical methods like DAGs)

The causal roadmap

1. Specify **estimand** (e.g. aided by target trial protocols)
2. Lay out assumptions needed for **identification** (e.g. aided by graphical methods like DAGs)
3. **Estimate** (e.g. by regression, matching, weighting, g-formula, doubly robust methods, machine learning, etc)

The causal roadmap

1. Specify **estimand** (e.g. aided by target trial protocols)
2. Lay out assumptions needed for **identification** (e.g. aided by graphical methods like DAGs)
3. **Estimate** (e.g. by regression, matching, weighting, g-formula, doubly robust methods, machine learning, etc)
4. **Sensitivity analysis** to assess assumptions

The ~~causal~~ roadmap

1. Specify **estimand** (e.g. aided by target trial protocols)
2. Lay out assumptions needed for **identification** (e.g. aided by graphical methods like DAGs)
3. **Estimate** (e.g. by regression, matching, weighting, g-formula, doubly robust methods, machine learning, etc)
4. **Sensitivity analysis** to assess assumptions

Can argue that this is a more general principe borrowing the logic of the scientific method

Example using clinical and nation-wide registers

- Data from **the Norwegian Trauma Registry** linked with various national registries on 26 562 trauma patients between 2015-2018 in the so-called NTR+ cohort dataset

Example using clinical and nation-wide registers

- Data from **the Norwegian Trauma Registry** linked with various national registries on 26 562 trauma patients between 2015-2018 in the so-called NTR+ cohort dataset

- How do we study the **effects of early opioid use** on long term use?

- **Effect of ”opioid saving” drug prescriptions (NSAIDs)** versus no NSAIDs week 3–10)

- **Effect of "opioid saving" drug prescriptions (NSAIDs)** versus no NSAIDs week 3–10)
- Can be thought of as **sustained strategies** $\bar{a} = \{a_0, a_1, a_2, ...\}$ assigned at baseline, where $a_l = 1$ if an individual still follow the assigned regime at day l and 0 otherwise

- **Effect of "opioid saving" drug prescriptions (NSAIDs)** versus no NSAIDs week 3–10)
- Can be thought of as **sustained strategies** $\bar{a} = \{a_0, a_1, a_2, ...\}$ assigned at baseline, where $a_l = 1$ if an individual still follow the assigned regime at day l and 0 otherwise

  Strategies followed is not identifiable for individuals at baseline

- **Effect of "opioid saving" drug prescriptions (NSAIDs)** versus no NSAIDs week 3–10)
- Can be thought of as **sustained strategies** $\bar{a} = \{a_0, a_1, a_2, ...\}$ assigned at baseline, where $a_l = 1$ if an individual still follow the assigned regime at day l and 0 otherwise

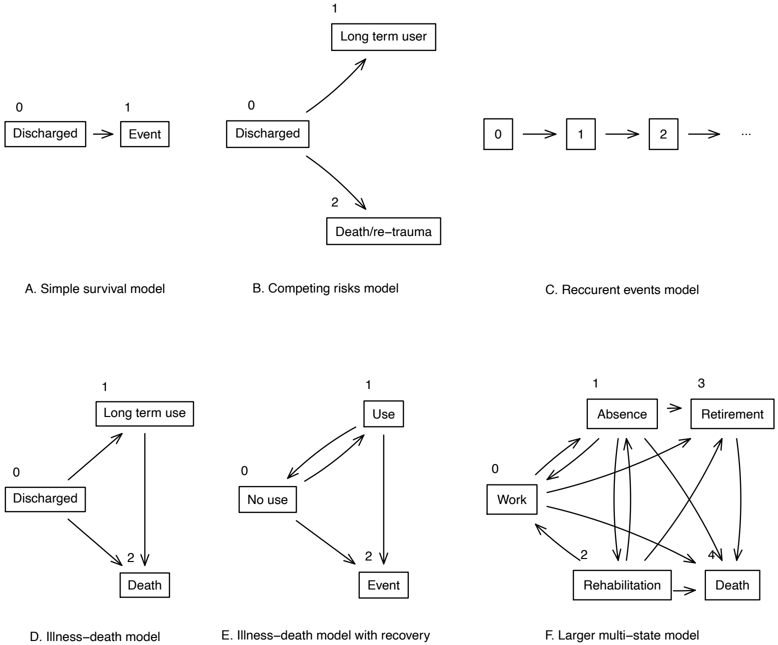
  Strategies followed is not identifiable for individuals at baseline
- Can use explicit target trial emulation using grace periods: **clone/censor/weight approach** (see e.g. Gaber et al. 2024)

- **Effect of "opioid saving" drug prescriptions (NSAIDs)** versus no NSAIDs week 3–10)
- Can be thought of as **sustained strategies** $\bar{a} = \{a_0, a_1, a_2, ...\}$ assigned at baseline, where $a_l = 1$ if an individual still follow the assigned regime at day l and 0 otherwise

  Strategies followed is not identifiable for individuals at baseline
- Can use explicit target trial emulation using grace periods: **clone/censor/weight approach** (see e.g. Gaber et al. 2024)

  Write out full protocol as in real RCT: specify eligibility criteria (opioid prescription day 0-14), time zero (14 days after discharge), treatment strategies (as above), baseline adjustment variables (age, sex, income, history of opioid use, geography, injury scores and type, comorbidities, hospital stay characteristics etc), time-varying adjustment variables (health visits in KUHR and NPR, opioids and other drug use) and long term opioid use as outcome (certain amount prescribed over 90 days, followed up by new prescriptions)

- **Broader outcomes of interest**:



Summed up

Summed up

- Formal causal language represent **a major advance** – bound to take over every area of research where intervention and treatment play a role

Summed up

- Formal causal language represent **a major advance** – bound to take over every area of research where intervention and treatment play a role
- Interventionism is about studying effects that are testable in principle; in other words it is about being **scientific**

Summed up

- Formal causal language represent **a major advance** – bound to take over every area of research where intervention and treatment play a role
- Interventionism is about studying effects that are testable in principle; in other words it is about being **scientific**
- Causal inference is not about being overly optimistic about causal claims, but **about being transparent and honest**