Causal inference in registry research

Magne Haugland Solheim
Øystein Ariansen Haaland

Copenhagen, January 2019

Health Registries for Research
Norway

NordForsk

Secret to a long life? Eating cake and sweets like Masazo Nonaka, world's oldest living man

Nonaka received the certificate from Guinness World Records in a ceremony at his home in Ashoro, on Japan's northern main island of Hokkaido, and celebrated the recognition with a big cake. Born on July 25, 1905, Nonaka grew up in a large family and succeeded his parents running the inn.

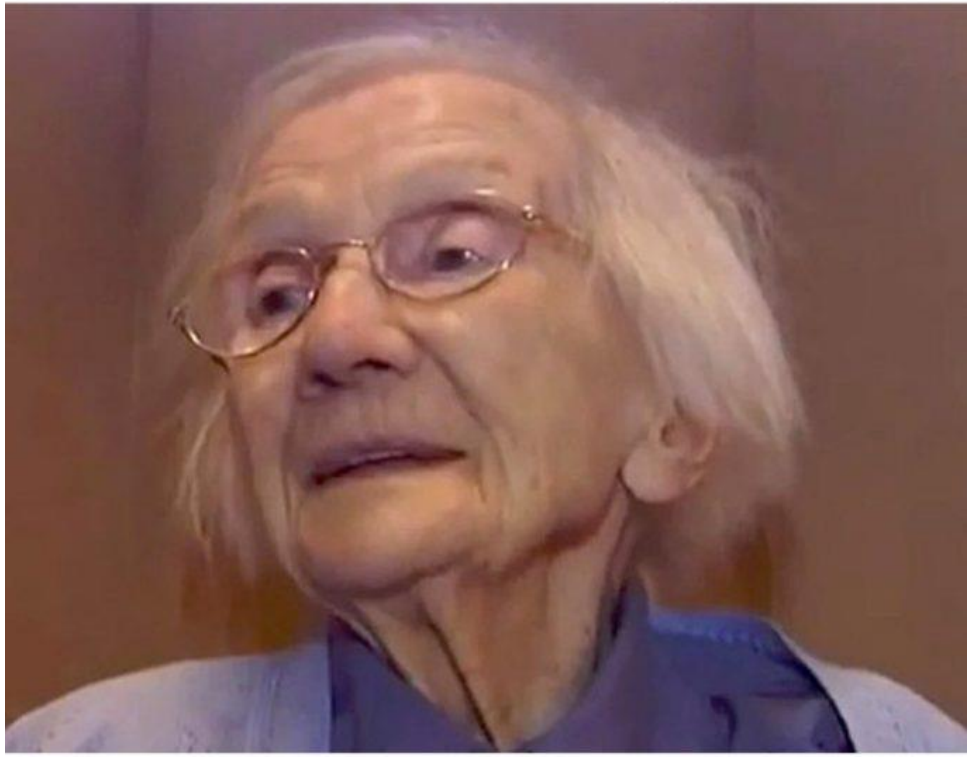
FITNESS Updated: Apr 10, 2018 17:23 IST

Associated Press, Tokyo

Japanese Masazo Nonaka, who was born 112 years and 259 days ago, eats his favourite cake as he receives a Guinness World Records certificate naming him the world's oldest man during a ceremony in Ashoro, Japan. (REUTERS)

Causal effects?

109-Year-Old Woman Says Avoiding Men Is The Secret To A Long Life



With 109 years of life, Jessie Gallan was officially the oldest woman in Scotland. Recently, she revealed the secret for longevity, which is surprisingly funny and simple.

read more:

Everybody knows what a causal effect is

- If only I had taken an aspirin before, I wouldn’t have a headache now
- Had he used a helmet, he would have survived the crash
- Had he not smoked, he would not have gotten lung cancer
- If I were younger, my back wouldn’t hurt

- Statements about what didn’t happen!

- We need to formalize the concept for scientific use and for clarifying under which conditions causal inference is possible

Rubin's causal model

- **The** framework for causal inference
- Potential outcomes and counterfactuals
- Example: flu and vaccine
  - Y: flu (= 0 if no flu, = 1 if flu)
  - T: vaccine (T = 0 if not vaccinated and T = 1 if vaccinated)

Each person has two potential outcomes: Y^1 if vaccinated and Y^0 if not.

One of the potential outcomes is observable, the other not. The unobserved outcome is called a counterfactual outcome.

What is the causal effect of vaccine on flu for one person?

- $E(Y^1_i - Y^0_i)$ individual causal effect for individual i
- Difference in outcome if he took the vaccine vs. if he didn't
- Problem: we cannot observe both Y^1_i and Y^0_i

  => Not possible to estimate individual causal effects

This is called *the fundamental problem of causal inference*

Missing data problem: causal inference is based on comparing counterfactual quantities that cannot be observed

Population causal effects can be estimated - sometimes

ATE = $E(Y^1 - Y^0)$ the average causal effect in the population (we can also use other measures like RR and OR).

- Comparing the situation when everyone is exposed with the case when no one is exposed
- Identification problem: No matter how much data we have, this cannot be estimated without further assumptions
- No causal inference without assumptions!
- Must find a set of assumptions that let us estimate ATE

Causal assumptions

- for estimating population causal effects

1. SUTVA (Stable Unit Treatment Value Assumption)

   - Non-interference: treatment assignment of one person does not affect potential outcomes of others (maybe not true for vaccine example?)

   - Only one version of the treatment/exposure

2. Positivity

   - Everyone has a positive chance of getting treated/exposed

3. **Ignorability** (The main issue)

   - Treatment/exposure assignment is independent of potential outcomes

   - often written like this: $Y^0, Y^1 \perp\!\!\!\perp T$ (some call it exchangeability)

Ignorability

- Allows us to connect potential (unobservable) outcomes to observed outcomes (i.e. data)
- Using the vaccination example:
- Ignorability of treatment assignment implies that the risk of getting a flu for those who did not get vaccinated if they (counterfactually) got vaccinated, would have been the same as for those who actually got the vaccine (and vice versa).
- In symbols: $E(Y^0|T=0) = E(Y^0|T=1) = E(Y^0)$ and $E(Y^1|T=0) = E(Y^1|T=1) = E(Y^1)$

so

$E(Y|T=0) = E(Y^0)$ and $E(Y|T=1) = E(Y^1)$ under ignorability

=> The observed risk of flu among the vaccinated is an unbiased causal estimate of the risk of getting the flu if everybody got the vaccine (and vice versa)

ATE under ignorability

$ATE = E(Y^1 - Y^0) = E(Y^1) – E(Y^0)$

Ignorability => $E(Y^1) = E(Y|T = 1)$ and $E(Y^0) = E(Y|T = 0)$

So we can estimate ATE by

- $\widehat{ATE} = E(Y|T = 1) – E(Y|T = 0)$

- If treatment assignment is ignorable: easy to estimate causal effects.
- Problem: ignorability is a very strong condition (we are basically saying that there is no confounding, selection etc.), and unlikely to hold without further steps
- Before we even consider estimating causal effects, we must justify that ignorability is reasonable in our study design

Identification strategies
- justifying ignorablitity

- **Identification comes before estimation!**

1. Randomization
2. Selection on observables
3. Others (will be covered later):
   - Natural experiments
   - Instrumental variables
   - Regression discontinuity and interrupted time series
   - Methods based on repeated measurements (fixed effects, DiD)

Randomization (RCTs)

- Randomization to treatment and control groups

  => Individuals are equal (in expectation) on all observed and unobserved variables
- Ignorability: if the groups were switched (before treatment), it would not affect the results
- The only thing that differs between groups is treatment
- Very credible identification strategy for causal effects
- RCTs often not possible, but

An ideal to aim for: Often helpful to imagine a RCT even when doing observational studies

Emulating RCTs

- even if you are doing an observational study

- If you had unlimited resources and no ethics: how could you set up a RCT to answer your research question?
- If you cannot imagine such an experiment (i.e. if you could not answer your question even under ideal conditions), the chances of answering the question with limited resources and poor data is pretty slim.
- Helps formulating research questions precisely
- No causation without manipulation (Holland)
- Can you for example think of a RCT to study the effect of age or gender on risk of stroke?

Emulating RCTs

- Old idea, at least since Cochran 1972
- Miguel Hernan (and co-workers) has written a lot about this.

Examples:

Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. American Journal of Epidemiology 2016; 183(8):758-764.

Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. Int J Obes (Lond). 2008 Aug;32 Suppl 3:S8-14.

Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. Journal of Clinical Epidemiology 2016; 79: 70-75.

Selection on observables

- Identification strategy: determine a set of variables to adjust for so that treatment assignment is conditionally independent of potential outcomes
- DAGs are great tools for this:

  - identify backdoor paths

  - don’t condition on colliders
- Conditional ignorability: potential outcomes are conditional independent of treatment assignment given X: $Y^0, Y^1 \perp\!\!\!\perp T|X$
- Alternative terms: no unmeasured confounding, Conditional Independence Assumption, no open backdoor paths, conditional exchangeability …
- Once the set X is determined: use regression, matching or weighting methods to adjust.

  => Causal effect estimates if assumption is OK.

Flu example

- Assume that elderly more often get vaccinated and that risk of flu also depends on age.
- Then ignorability fails: the risk of flu for unvaccinated if they counterfactually actually took the vaccine, would not be the same as for those that actually got vaccinated.
- If age is the only confounder (highly unlikely!), if we condition on age, we achieve conditional ignorability. For each age the risk of flu for unvaccinated if they counterfactually actually took the vaccine, would be the same as for those that actually got vaccinated.

=> As if vaccine was randomly distributed (with different probabilities) within each age group

Selection on observables

= no unmeasured confounding

Selection on observables is a very strong assumption, and require data on all covariates you need to adjust for.

- Basically we are assuming a RCT within each “stratum” of covariates
  (for individuals with the same value of X, exposure is random)

- Have we found the correct set X? (Unverifiable with observed data)
- Do we have data on all X?
- Health registries: data not collected for research purposes
  => Selection on observables is often a dubious assumption

Methods of estimation

-secondary, will not cover this in detail

- Identification comes before estimation
- Once you have decided on an identification strategy, the next step is to estimate the causal effect of exposure on outcome.
- Many options: regression, matching or weighting
- Each have strengths and weaknesses, but properly done, they should give similar results. (Exception is time varying confounding, where weighting methods works best)
- Notice: Identification is key. If your identification strategy sucks, no sophisticated estimation method can save you.
- Sometimes misunderstood: choosing for example matching instead of regression does not make your estimates more causal