Causal inference in pediatric and perinatal epidemiologic research: *From questions to methods*

Jessica Young

June 15, 2021

We often care about causal questions

Many (if not the majority) of questions studied in pediatric and perinatal epidemiologic research are causal questions.

- What is a “causal question”?
- How does “causal inference” differ from “statistical inference”?

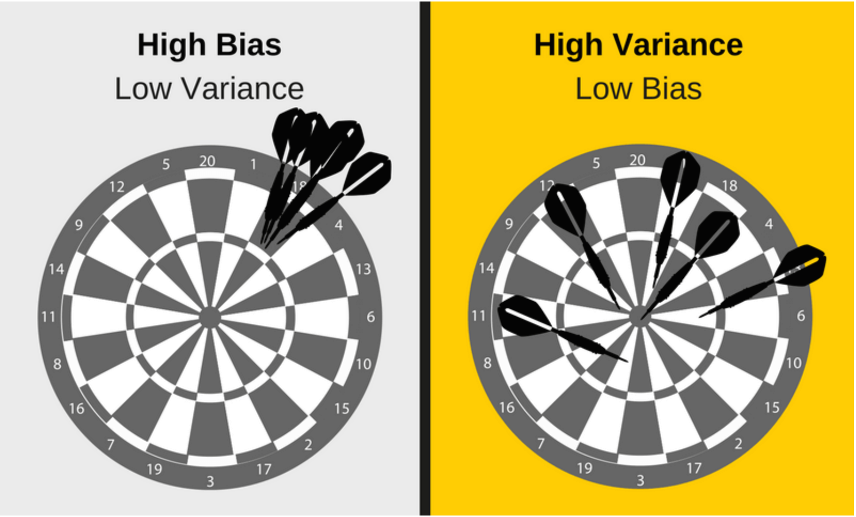
Statistical Inference

Statistics are simply computational algorithms applied to a set of measurements (data)

- Popular examples: sample average, sample variance, sample average difference comparing two groups

Statistical inference is the process of using data from a sample to learn about a particular feature of a population

Can think of any dart as a statistic computed from data in one sample



Want the darts to scatter equally around the center (unbiased).
Want tight scatter around the center (low variance)

Center of the dart board

The center represents the population feature we want the statistic (the data) to tell us something about: the question we want to answer

- **However statistical machinery is limited to questions about population features we can, at least in principle, observe (or measure)**
- **Call these "statistical parameters"**

Examples of "statistical parameters":

- In a population of 5 year olds, the mean of a cognitive score at age 10
- The mean difference in this cognitive score at age 10 among subset of these who initiated a particular medication at age 5 versus did not initiate at age 5.

Problem: many research questions are about causal effects which are not statistical parameters.

What is a causal effect?

Contrast (e.g. difference) of outcomes in the same individuals but under different treatments.

- Ex: For this same population of 5 year olds, the mean difference in age 10 cognitive score had they all initiated the med at 5 versus, instead, had they all not initiated.
- This is a causal effect (of initiating this medication age 5 on mean cog score at age 10) because any difference MUST be due to the *treatment*.

Causal effects contrast *counterfactual* (*potential*) features of a population.

- For any individual, cognitive score can only be observed under, at most, one potential treatments.

Rather than statistical parameters, causal effects are counterfactual parameters.

Causal inference versus Statistical inference

When interest in a causal effect, extra steps needed to choose a statistic (approach to analyzing our data)

1. Articulate the causal effect we want
2. Consider subject matter assumptions that let us equate this effect to some statistical parameter, function of only measured characteristics – *identifying* assumptions
   - ex: "no unmeasured confounding"
3. Finally we can choose a statistic for that statistical parameter and understand its "dart board" properties. Different statistics can be justified under different additional assumptions after fixing steps 1 and 2
   - *statistical assumptions* – ex: outcome is normally distributed conditional on treatment status

Step 3 constitutes statistical inference. Adding steps 1 and 2 constitutes causal inference.

Overview

Part I: Articulating causal questions

- Notation: factuals versus counterfactuals
- Examples of statistical (factual) versus counterfactual parameters
- Review of key types of causal effects
  - Marginal versus conditional effects
  - Time-fixed versus time-varying treatment effects
  - Effects of static versus dynamic treatment rules
  - Effects of deterministic versus stochastic treatment rules
- *Target trials* as an aid for question articulation

Part II: Identification: Linking counterfactual parameters to statistical parameters

- Causal inference when we don't have an "idealized" version of target trial
- Key *identifying* assumptions and causal diagrams
- The g-formula – a core statistical parameter in causal inference

Part III: Statistics

- The many representations of the g-formula and how this connects to statistics
- Inverse probability weighting
- Marginal structural models

Part IV: When counterfactual contrasts are not causal effects

- The problem of conditioning on post-treatment variables

Part I: Articulating causal questions

Notation: Factuals and statistical parameters

Consider a population (e.g. 5 year olds). Returning to the previous example, define for each child in this population

- A an indicator of whether a medication of interest initiated at age 5.
- Y cognitive score at age 10

A and Y here are nothing more than characteristics that we can in principle measure on children from this population.

Examples of statistical parameters

Generically using the notation $E[Y]$ to represent the “mean of Y in the population”, some examples of statistical parameters with respect to these characteristics are

- $E[Y]$, the overall (marginal) population mean score
- $E[Y|A=1]$ the population mean score among those who initiated (i.e. conditional on initiating)
- $E[Y|A=1] - E[Y|A=0]$ the population mean difference among those who initiated versus among those who did not

And there are more examples. Well established literature on statistics (methods) for targeting these statistical parameters.

Association $\neq$ causation...except when it is

We are taught early on that a difference like $E[Y|A=1]-E[Y|A=0]$ quantifies "association not causation". This is consistent with its failure to meet our definition of a causal effect

- $E[Y|A=1]-E[Y|A=0]$ does not contrast outcomes under different treatments in the same individuals
- Instead it compares outcomes in DIFFERENT individuals experiencing different treatment scenarios.

However, it might be possible to equate this statistical parameter to some causal effect under plausible assumptions (Part II). To do this we have to have a way to define a causal effect as formally as we have defined this statistical parameter.

Counterfactual outcomes

Define

- $Y^{a=1}$ as the score a child would have experienced at age 10 had, possibly contrary to fact, they initiated the medication at age 5.
- $Y^{a=0}$ similarly is this score had instead they not initiated at this time.

In turn we can formally write the causal effect of initiating (versus not) on mean cognitive score at age 10 in this population as (on the additive scale):

$$E[Y^{a=1}] - E[Y^{a=0}]$$

Sometimes call this an "average treatment effect" – average of individual level effects $Y^{a=1} - Y^{a=0}$.

Marginal versus conditional effects

The causal effect

$$E[Y^{a=1}] - E[Y^{a=0}]$$

is an example of an *overall* or *marginal* effect in that it refers to the entire study population. In some instances we might be interested in a *conditional* effect, within subsets of this original population, e.g.

$$E[Y^{a=1}|\text{started reading by age 4}] - E[Y^{a=0}|\text{started reading by age 4}]$$

the effect among the subset of children who started reading by age 4.

- Still meets our definition of a causal effect.

Time-fixed versus time-varying treatment effects

- Our running example is a case of a *time-fixed* (or *point*) treatment effect (initiating versus not at age 5)
- Many questions in pediatric and perinatal epidemiology are about time-varying treatment effects.
- Clarifying the distinction will force us to be more precise about the role of time in our question.

Time-fixed example

Considered the average (overall) causal effect of initiating a med at age 5 (versus not) on a score at age 10: The causal effect

$$E[Y^{a=1}] - E[Y^{a=0}]$$

Thus far we have been somewhat vague because “age 5” or “age 10” can mean the day the child turns that age or the day before they turn the next age, or lots in between. Let’s make this question more precise and suppose we mean the child’s 5th and 10th birthdays.

Time-varying example

Naturally, we might not only be interested in the effect of initiating this medication or not at 5 but maybe patterns of use or even timing of initiation between age 5 and age 10. To formally accommodate this case, let's define a broader, and more precise, set of (factual) characteristics of children in this study population. Define:

- Y as cognitive score when child turns 10
- A_t as an indicator of whether the child took the medication on day t

where $t = 0, \ldots, K$ and $t = 0$ refers to the day the child turned 5, K indexes 5×365 later.

Time-varying example

Now we are conceptualizing not only (factual) characteristics at the start and end of a follow-up period of interest but time-evolving characteristics during this period as well. Some notation:

- Use overline notation to denote history e.g. $\overline{A}_t = (A_0, A_1, \ldots, A_t)$
- Use underline notation to denote future e.g. $\underline{A}_t = (A_t, A_{t+1}, \ldots, A_K)$
- Sometimes I'll drop the t subscript if it's clear which we mean.

Can define a wide variety of causal effects on this time-varying process because there are many, many rules we can consider to intervene on this process.

"Always" versus "never" treat

For example, could consider the effect of ensuring all children in the population take medication on every day $t = 0, \ldots, K$ versus never take it. Typically write as:

$$E[Y^{\overline{a}=\overline{1}}] - E[Y^{\overline{a}=\overline{0}}]$$

where $\overline{a}$ indexes the outcome under "Set $\overline{A}_K$ to some instantiation $\overline{a}_K \equiv \overline{a}$".

- We call these types of "treatment rules" *static* rules because the treatment assignment under the rule at every future time is known at the start (does not depend on what happens to the child over time).

Another example

Another example of a static causal effect: could consider the effect of ensuring all children take medication consistently for the first year and then stop for the remaining 4 years, versus never take it. Can write this as

$$E[Y^{\bar{a}_{365}=\bar{1}_{365},\underline{a}_{366}=\underline{0}_{366}}] - E[Y^{\bar{a}=\bar{0}}]$$

This is one possible way to formalize a question about "sensitivity periods" often often of interest in studies of early life exposures.

Static rules

While static rules may be of interest in many settings, they are problematic in two (linked) ways:

- Static rules are often unethical and/or unrealistic– e.g.
  - a child that doesn't take treatment at the start may develop a condition indicating it later.
  - a child that initially take treatment may develop a contraindication for it (or not need it anymore)
- This makes questions about causal effects of static rules harder to answer using real-world data (Part II)

Dynamic rules

Dynamic rules more realistically account for time-evolving circumstances that relate to treatment decisions. E.g. Average (overall) causal effect on outcome under the different rules (applied on all days $t = 0, \ldots, K$)

- Take medication on day t if no contraindication has developed by that day; otherwise, do not take medication on day t
- Do not take medication on day t if no indication has developed by that day; otherwise take medication on day t

However, in some cases we may still want even more flexibility to define the question of interest...

Deterministic versus Stochastic rules

The examples we've considered so far are cases of so-called *deterministic* treatment rules

- Meaning that the rule is stated in such a way that, at every time t, either all individuals, or individuals with a common characteristic, are guaranteed the same level of treatment.
- In some cases, it can be quite difficult to come up with realistic deterministic rules even when they are dynamic.
- This is particularly the case when we consider continuous treatments (exposures), or otherwise treatments that take on many levels in the real (factual) world.

Example: Questions about effects of exercise interventions

Consider interest in the causal effects of interventions on daily minutes of exercise over time. Many ways to define an average (overall) causal effect in this case depending on the interventions we specify:

- Static (deterministic) example: "Everyone exercise exactly 30 minutes every day over the next 5 years"
- Dynamic (deterministic) example: "Exercise exactly 90 minutes on each day t that you have normal blood pressure; otherwise exercise 30 minutes" on all days over next five years.

While second is perhaps more realistic than first, estimating average outcomes under either of these rules will rely on particularly strong assumptions when there is wide variation in the patterns of the number of minutes that people exercise in reality (Part II). And they are too specific to be useful from a policy perspective.

Representative interventions

What about the average causal effect of “Ensuring a distribution such that everyone exercise *at least* 30 minutes per day” versus “less than 30 minutes”? or even compared to “no intervention”?

- This is more realistic, allowing potentially wide variation in treatment levels under the rule, even for people with similar characteristics.
- However, it’s vague in many respects including
  - It doesn’t specify what the distribution of treatment would be under this rule.
  - There are an infinite number of distributions consistent with how we stated this.
- There is one distributional choice that is especially convenient in that it leads us to a statistical parameter that is more tractable to estimate than other choices.
  - Representative interventions.

Generalized counterfactual notation

More generally, for g any treatment rule of interest,

- Define Y^g as the outcome for an individual in the population had, possibly contrary to fact, they adhered to the treatment rule g

For two choices of g – denote these g_1 and g_2 – then we can define the average (overall) causal effect of adhering to g_1 vs. g_2 on the outcome as

$$E[Y^{g_1}] - E[Y^{g_2}]$$

What about study design?

- Thus far, we have been discussing how to articulate causal *questions*
- We have been agnostic as to whether a randomized or nonrandomized (observational) study would be used to try to *answer* that question.
- Regardless of what design is used, assumptions will *always* be needed to link causal (counterfactual) parameters to statistical (factual) parameters.
- Randomized studies have been historically equated with causal inference because, *in some cases*, the assumptions we need to make this link are guaranteed, at least in principle, by the design.

The target trial

The concept of a target trial has been growing in popularity as a way to more transparently communicate interest in estimating a causal effect when using data collected in a nonrandomized (observational study).

- This is the trial that the investigator would have implemented in order to answer their underlying causal effect of interest.
- This has turned out to be a somewhat more palatable way to the scientific community to allow investigators to admit interest in causal (counterfactual) questions without using counterfactual language.

Uphill Battle

Even articulating causal questions through this target trial analogy remains a bit of an uphill battle given that, for more than a century, powerful voices in science (particularly early "fathers" of statistics) made explicit discussion of causality taboo outside of randomized studies.

- History of the "divorce" between statistics and causality (Judea Pearl, *Book of Why*)
- *The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data*, AJPH, Hernán

Target trial protocol

Writing out the protocol of the target trial forces the investigator to think about and articulate the components of their causal question explicitly. For example

- That is the time of randomization? This is our $t = 0$ in our formal conceptualization (time 0).
- Who is eligible for the trial (the population that the expectation refers to in our counterfactual contrast)
- What are the outcomes, when would they be measured.
- What are the treatment rules? g_1, g_2 (or more if multiple "arms")

“Emulating” a target trial

The task of estimating the causal effect of interest implied by this protocol in an observational study is popularly referred to as “emulating” the target trial with the observational data. This may include for example

- making sure that the follow-up information that is used for each individual in the analytic data set begins at the time we would randomize in the imagined target trial.
- Hernán and Robins, *Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available*, AJE, 2016

In Part II we are going to be more formal about what this “emulation” really means.

Questions?

Part II: Identification of causal effects with real-world (factual) data

“Idealized” version of a target trial

Imagine we were able to conduct a randomized trial where individuals meeting eligibility for the study population are enrolled and then randomized to say one of two study arms where the protocol is to follow different rules of interest (g_1 or g_2, e.g. on time-varying medication use from 5 to 10). Also assume

1. Everyone adheres to the protocol for the study duration (e.g. 5-years)
2. The outcome (e.g. cognitive score) at the end of the study is known for everyone enrolled and randomized at the start.

Randomization and identification

Let Z represent the result of a coin toss ($Z = 1$ or heads you are assigned arm 1, $Z = 0$ or tails, arm 2). Because the value of Z is determined only by the coin flip, we have that

$$Y^g \coprod Z$$

where $\coprod$ denotes independence for either $g = g_1$ or $g = g_2$. In other words, the value of Z an individual gets is not associated with their future counterfactual outcomes.

- This independence is an assumption but it is guaranteed to hold by the definition of Z.

Idealized study

This study design, which guarantees this independence assumption, PLUS the idealized nature of the trail (perfect adherence and complete outcome measurement), allows us to equate the causal effect $E[Y^{g_1}] - E[Y^{g_2}]$ to the statistical parameter

$$E[Y|Z = 1] - E[Y|Z = 0]$$

Alternatively we say the effect we want is *identified* by this statistical parameter.

Idealized study

Question: How would you analyze the data of this trial (where we've measured Z and Y for everyone enrolled) given we've now established our statistical parameter is:

$$E[Y|Z=1]-E[Y|Z=0]?$$

Remembering that Y is just the outcome of interest, Z indicates treatment arm.

Real world study

Real world studies are rarely "ideal" particularly for causal effects of time-varying treatments.

- People don't follow the protocol
- People drop out of the study (*censoring* of outcomes)
- A trial may not be feasible or timely, we may have data such that nothing is randomized (investigator has no control at any time in determining what treatment people might take, there is no coin flipping, no Z) – observational study

Also people sometimes die (or experience some other event) that makes the outcome we care about impossible (or meaningless – see shared readings (Chiu et al, Snowden et al, Young & Stensrud).

Real world study

We still have causal questions even though we rarely have idealized trials to answer them. In this more realistic case:

- To equate our causal effect to a statistical parameter, we'll have to make assumptions that are not guaranteed.
- Further the statistical parameter generally more complex than a simple comparison of outcome means.

Observational study

Suppose we conduct an observational study where the following longitudinal data was collected on each of n individuals meeting criteria for the study population of interest at time 0.

- Y cognitive score when child turns 10
- In every interval (e.g. day/week/year) $t = 0, \ldots, K$, we measure
  - A_t an indicator of whether the child took the medication on day t
  - L_t a vector of covariate measurements in interval t (other medications, newly diagnosed diseases, behaviors)

  where L_0 includes pre time 0 characteristics possibly child sex, parental health history, health conditions prior to age 5, birthweight.

There is no Z (coin) in this study. The investigators simply observe, they don't flip coins and assign anything.

Task

Now consider assumptions that let us link our causal effect of interest $\mathrm{E}[Y^{g_1}] - \mathrm{E}[Y^{g_2}]$ to a statistical parameter, that is some function of the measured variables.

- This is the same question we had in the case of the idealized trial where all we needed for unbiased statistical analysis were measures of (Z, Y).
- But the change in study designs means we need more complex assumptions to link this same effect to a function of the measured study variables $(\overline{L}_K, \overline{A}_K, Y)$

Exchangeability

Consider an assumption very similar to the counterfactual independence that was guaranteed in the ideal trial but modified:

$$Y^g \coprod A_t | \overline{L}_t, \overline{A}_{t-1} = \overline{a}^g_{t-1}$$

This states that the counterfactual outcome under rule g is independent of the treatment actually received on a given day t within levels of measured covariate history.

- Many names for this assumption including *exchangeability* and *no unmeasured confounding*
- Can refer to the covariate history $\overline{L}_t$ as the *measured confounder history*
- This independence guaranteed in a *sequentially randomized trial* where A_t assigned by weighted coin, at most dependent on measured past (not guaranteed in observational study)

Causal graphs to evaluate exchangeability

Causal graphs are a way to explicitly represent subject matter assumptions/background knowledge that would

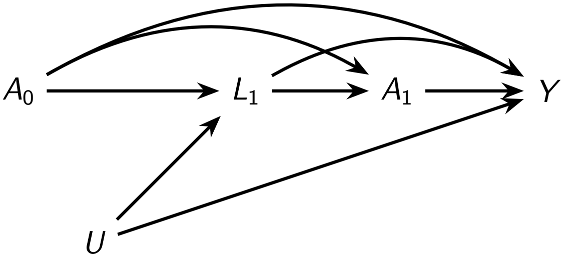
- support (or fail to support) an exchangeability assumption (this counterfactual independence)

Causal Directed Acyclic Graphs (DAGs)

Causal DAGs in particular depict the underlying causal structure of statistical dependence between variables:

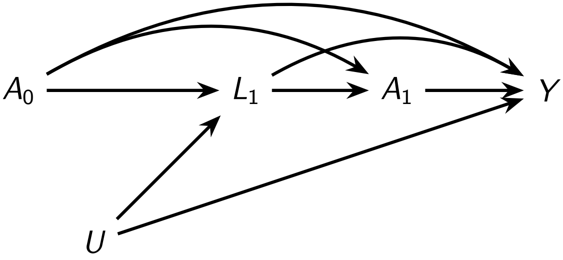
- They can be used to represent an underlying *nonparametric counterfactual causal model*
- They don't require assumptions about the functional form of dependence (e.g. linear, quadratic) or distributional assumptions (Normal, Poisson).
- The only assumptions they make on the nature of the dependence between any two temporally ordered variables X and Y on the causal DAG are
  - Can X cause Y? – If there is no path consisting of directed arrows connecting X and Y assumes answer is no
  - Do X and Y share common causes? – If no common cause depicted, assumes answer is no

Example of a causal DAG for our data structure (2 timepoints, $K = 1$), conditioned on level of L_0



Basics: Paths connecting two nodes with directed arrows represent causal structure of dependence (e.g. $A_0 \rightarrow L_1$). Backdoor paths represent noncausal structure (e.g. $A_1 \leftarrow L_1 \leftarrow U \rightarrow Y$).

Sequential randomization assumption



Allows that measured time-varying covariates are affected by past treatment. Also that there could be unmeasured common causes U of these measured covariates and the outcome but they do not directly affect treatment (by no arrow from U to A_0 or A_1).

Sequential randomization assumption

- This assumes that the underlying data generating process of the observational data is that of a sequentially randomized study
  - Where A_t at each time is at most "assigned" by past measured covariates
  - Even though the study investigators didn't do the "assigning" (because this is an observational study)
- Under a data generating assumption like this (sequential randomization, no arrows from U into A_t for any t) then exchangeability

$$Y^g \coprod A_t | \overline{L}_t, \overline{A}_{t-1} = \overline{a}^g_{t-1}$$

holds for any choice of rule g

Sufficient but not necessary

Sequential randomization is a sufficient condition but it isn't a necessary condition.

- More generally, the rules to evaluate exchangeability for a rule g and a particular set of measured covariates $\overline{L}_t$ at each t from a causal diagram are more involved.
- They are also dependent on the choice of rule g
- For the same treatment (e.g. medication in our example), there may be a data generating mechanism such that exchangeability holds for a static rule g assigning this treatment but fails for a dynamic rule g
- Depends on the types of assumptions we make about unmeasured common causes

SWIGs

The best way to reason about this generally is using Single World Intervention Graphs (SWIGs) – Richardson and Robins, 2013

- A SWIG is specific to a particular counterfactual "world" in which g is implemented
- Transformation of the causal DAG that explicitly depicts counterfactuals under g
- Have added some additional slides with examples at end.

Consistency

Consistency: If individual has treatment history consistent with g then future covariates and outcome are the values they would take under g for all $t = 0, \ldots, K$

- If $\overline{A}_t = \overline{a}^g_t$ then $\overline{L}^g_{t+1} = \overline{L}_{t+1}$, $Y^g = Y$.
- Requires there are not "multiple versions of treatment"
- Allows us to link counterfactuals to factuals
- Another assumptions required for this linkage is "no interference" (other people's treatments do not affect my counterfactual outcomes) – both part of so-called Stable Unit Treatment Value Assumption (SUTVA) – Rubin

Consistency

- Violations of “no interference” are hard to deal with – fundamental issue in studies of infectious disease as well as interventions on social behaviors
- We can always avoid violations of consistency by being as clear as possible about what g we are thinking about
- Often this requires us to be more explicit about what the real underlying treatment/exposure is (the thing we would intervene on if we could across different counterfactual “worlds”)
- It might be something complicated or something we didn’t even measure

Back to exercise

Example: for A_t minutes of exercise in interval t

- Define $R_t = I(A_t > 30)$, that is an indicator that an individual exercised at least 30 minutes in interval t
- We considered interventions that would ensure "R_t is always set to 1" for all t
- Had we defined our counterfactual outcomes with respect to this R_t intervention, consistency not reasonable
- We avoided this by defining counterfactual outcomes in terms of more explicit intervention rules on A_t with this "at least" property.

Representative interventions

However, it turns out that picking representative intervention on A_t as a more precise statement of this question let's us use R_t in place of A_t in certain algorithms, which leads to more tractable computation (Part III). But A_t, not R_t, is the conceptual exposure.

- This is due to the fact R_t is a *coarsening* of A_t
- Separates/groups the support of A_t (values it can take)
- Stitelman et al. *The impact of coarsening the explanatory variable of interest in making causal inferences: Implicit assumptions behind dichotomizing variables* http://biostats.bepress.com/ucbbiostat/paper264, 2010.
- Vanderweele and Hernán. *Causal Inference Under Multiple Versions of Treatment.* JCI, 2013.
- Young et al. *Inverse probability weighted estimation of risk under representative interventions in observational studies*, JASA 2019

Representative interventions

This idea can in principle be extended to overcome consistency violations in other settings but at the expense of a coarsening *assumption* that isn't guaranteed as it is in our exercise example.

- E.g. "Effect of BMI" is ill-defined
- investigators asking this question may have more well-defined questions in mind, but typically real exposures are high-dimensional, even unmeasured
- Under *assumption* that BMI is a *coarsening* of those exposures, might analyze data with BMI in role of exposure but conceptually it isn't – the question is about effects of interventions on these unmeasured exposures.
- Call these interventions *proxy representative* interventions – with BMI serving as a proxy in the algorithm for the real exposures.
- Aris et al. *Separating Algorithms from Questions and Causal Inference with Unmeasured Exposures: An Application to Birth Cohort Studies of Early BMI Rebound*, AJE 2021

Summary thoughts on consistency

- Consistency violations tend to arise when we call something "exposure" that will "act" as exposure in a statistical analysis but doesn't coincide with the conceptual exposure – what we would intervene on if forced to be explicit
- We can always avoid this by not conflating algorithms with questions
- This solves consistency but may bring to light that we need extra assumptions beyond the well-established assumptions of exchangeability, consistency, and postivity (discussed next) to answer our question.

Positivity

Positivity: For any measured confounder and treatment history plausible in the observational study and consistent with g prior to time t, it must be possible to observe a value of treatment consistent with g at time t, for all times t.

- that is, your real-world data must "support" a question about g

Like exchangeability, depends on choice of g:

- E.g. if g is "always treat" and past measured confounder history includes contraindication for treatment, positivity will be violated. Could rectify this by changing to a dynamic rule that accommodates this reality.

That is, you can avoid positivity violations by modifying your question. Stochastic rules may be required to avoid this for continuous treatments.

The g-formula

Robins (1986,1987) proved that, given exchangeability, consistency, positivity for a general time-varying treatment rule g, can identify $\mathrm{E}[Y^g]$ by the statistical parameter:

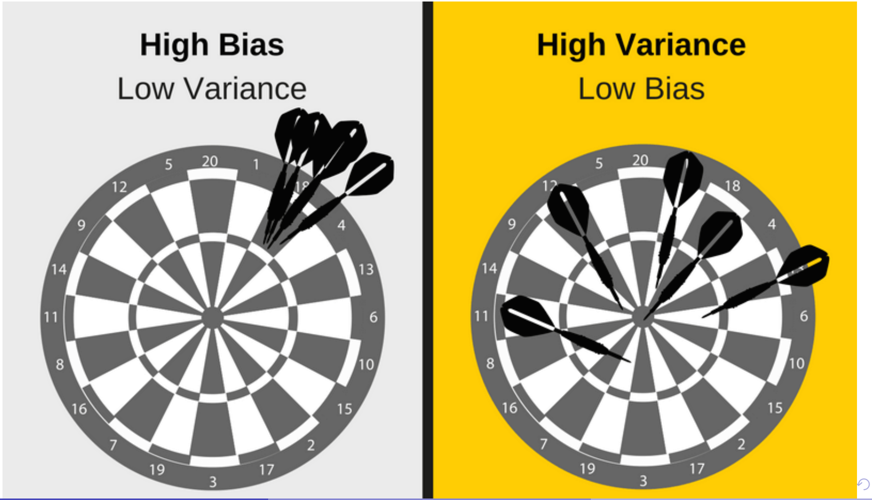
$$\sum_{\bar{a}_K,\bar{l}_K} \mathrm{E}[Y|\overline{A}_K = \bar{a}_K, \overline{L}_K = \bar{l}_K] \prod_{t=0}^{K} f^g(a_t|\bar{a}_{t-1}, \bar{l}_t) \prod_{t=0}^{K} f(l_t|\bar{a}_{t-1}, \bar{l}_{t-1})$$

Referred to this as the *g-computation algorithm formula* indexed by rule g. Modern literature shortened to *g-formula*. In turn a difference in this formula indexed by two different rules g_1 versus g_2 may identify the causal effect

$$E[Y^{g_1}] - E[Y^{g_2}]$$

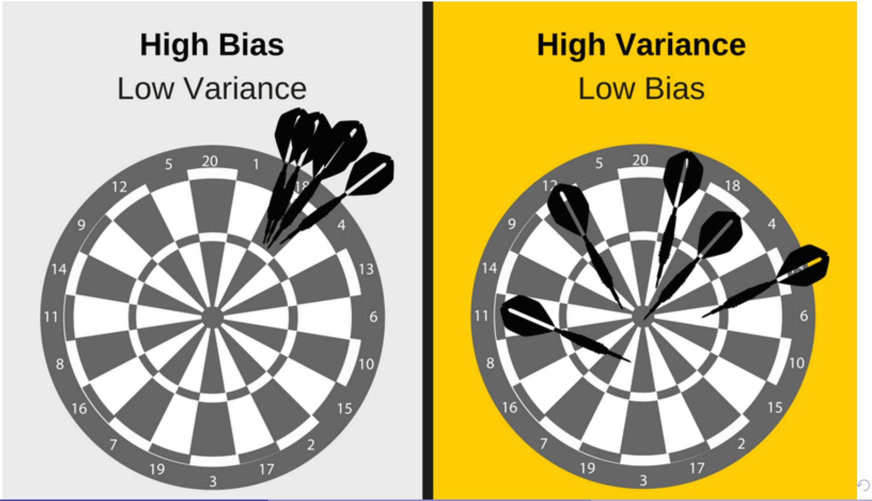
Back to the dart board!

Now we have what we need to employ statistical machinery! We know what the center is (a contrast in g-formulas for different choices of g).



Back to the dart board!

Our challenge is that this is much nastier than the statistical parameters we're used to and so the statistics we need to construct are generally more complicated.



Understanding the g-formula

Before we get to this, let's understand this formula a bit better – it's pieces, as well as some special cases.

$$\sum_{\bar{a}_K, \bar{l}_K} \mathrm{E}[Y | \overline{A}_K = \bar{a}_K, \overline{L}_K = \bar{l}_K] \prod_{t=0}^{K} f^g(a_t | \bar{a}_{t-1}, \bar{l}_t) \prod_{t=0}^{K} f(l_t | \bar{a}_{t-1}, \bar{l}_{t-1})$$

The symbol $\sum$ means sum and $\prod$ product.

- The g-formula in general is a sum over all possible levels of the time-varying treatment and measured confounder histories.
- Can think of enumerating all the possible levels of these histories, calculating the product inside for each one, and then summing.
- Generally not feasible in most settings – too many to enumerate.
- If any component of history is truly continuous, impossible. And need to replace sum with integral.

Understanding the g-formula

Let's now understand the pieces inside the sum.

$$\sum_{\overline{a}_K, \overline{l}_K} \mathrm{E}[Y | \overline{A}_K = \overline{a}_K, \overline{L}_K = \overline{l}_K] \prod_{t=0}^{K} f^g(a_t | \overline{a}_{t-1}, \overline{l}_t) \prod_{t=0}^{K} f(l_t | \overline{a}_{t-1}, \overline{l}_{t-1})$$

- The term $\mathrm{E}[Y | \overline{A}_K = \overline{a}_K, \overline{L}_K = \overline{l}_K]$ is the (factual) outcome mean conditional on a particular treatment and confounder history – this quantity should be very familiar!
- The term $f(l_t | \overline{a}_{t-1}, \overline{l}_{t-1})$ (thinking discretely) is the (factual) chance of having level l_t of the assumed measured confounders at t conditional on having the particular history $(\overline{a}_{t-1}, \overline{l}_{t-1})$.
- $\prod_{t=0}^{K} f(l_t | \overline{a}_{t-1}, \overline{l}_{t-1})$ means we are taking the product of this chance over all times $t = 0, \ldots, K$.

Understanding the g-formula

Finally, the part of this formula that is specific to the rule g is the term $f^g(a_t|\overline{a}_{t-1}, \overline{l}_t)$. I like to call this the *intervention treatment distribution* associated with g.

$$\sum_{\overline{a}_K, \overline{l}_K} \mathrm{E}[Y|\overline{A}_K = \overline{a}_K, \overline{L}_K = \overline{l}_K] \prod_{t=0}^{K} f^g(a_t|\overline{a}_{t-1}, \overline{l}_t) \prod_{t=0}^{K} f(l_t|\overline{a}_{t-1}, \overline{l}_{t-1})$$

- This is the chance of receiving the particular treatment level a_t at time t conditional on having the particular history level $(\overline{a}_{t-1}, \overline{l}_t)$ according to the rule g
- While the other pieces of the g-formula are factual features of the population, this piece is a feature of our question – although in some special cases, it may *also* depend on features of the factual population.

Example: "Always Treat"

Suppose we choose g to be the static rule "Always treat", that is "Set $A_t = 1$ at all $t = 0, \ldots, K$. In this special, case the g-formula expression

$$\sum_{\bar{a}_K, \bar{l}_K} \mathrm{E}[Y|\bar{A}_K = \bar{a}_K, \bar{L}_K = \bar{l}_K] \prod_{t=0}^{K} f^g(a_t|\bar{a}_{t-1}, \bar{l}_t) \prod_{t=0}^{K} f(l_t|\bar{a}_{t-1}, \bar{l}_{t-1})$$

reduces to

$$\sum_{\bar{l}_K} \mathrm{E}[Y|\bar{A}_K = \bar{1}_K, \bar{L}_K = \bar{l}_K] \prod_{t=0}^{K} f(l_t|\bar{1}_{t-1}, \bar{l}_{t-1})$$

This is because under this rule $f^g(a_t|\bar{a}_{t-1}, \bar{l}_t) = I(a_t = 1)$; the chance under g (for any confounder history) of receiving the treatment is 1 and the chance of not receiving the treatment is 0.

The g-formula for generic (deterministic) static rules

By same logic, for any static rule g “Set $A_t = a^*_t$ for some selected constant a^*_t, $t = 0, \ldots, K$ the g-formula expression

$$\sum_{\bar{a}_K, \bar{l}_K} \mathrm{E}[Y|\overline{A}_K = \bar{a}_K, \overline{L}_K = \bar{l}_K] \prod_{t=0}^{K} f^g(a_t|\bar{a}_{t-1}, \bar{l}_t) \prod_{t=0}^{K} f(l_t|\bar{a}_{t-1}, \bar{l}_t)$$

reduces to

$$\sum_{\bar{l}_K} \mathrm{E}[Y|\overline{A}_K = \overline{a^*}_K, \overline{L}_K = \bar{l}_K] \prod_{t=0}^{K} f(l_t|\overline{a^*}_{t-1}, \bar{l}_{t-1})$$

The g-formula for generic (deterministic) dynamic rules

When indexed by a deterministic dynamic rule g, the g-formula expression

$$\sum_{\overline{a}_K,\bar{l}_K} \mathrm{E}[Y|\overline{A}_K = \overline{a}_K, \overline{L}_K = \bar{l}_K] \prod_{t=0}^{K} f^g(a_t|\overline{a}_{t-1}, \bar{l}_t) \prod_{t=0}^{K} f(l_t|\overline{a}_{t-1}, \bar{l}_t)$$

reduces to

$$\sum_{\bar{l}_K} \mathrm{E}[Y|\overline{A}_K = \overline{a}^g_{K-1}, \overline{L}_K = \bar{l}_K] \prod_{t=0}^{K} f(l_t|\overline{a}^g_{t-1}, \bar{l}_t)$$

where, unlike for static rules, a^g_t is only a prespecified constant within particular levels of the confounder history.

The g-formula for stochastic rules

This clarifies that the general form of the g-formula

$$\sum_{\overline{a}_K,\overline{l}_K} \mathrm{E}[Y|\overline{A}_K = \overline{a}_K, \overline{L}_K = \overline{l}_K] \prod_{t=0}^{K} f^{g}(a_t|\overline{a}_{t-1}, \overline{l}_t) \prod_{t=0}^{K} f(l_t|\overline{a}_{t-1}, \overline{l}_{t-1})$$

only needed for “stochastic rules” such that

- the intervention treatment distribution is not *degenerate*
- the chance of receiving some level of treatment under g takes values between 0 and 1 for some levels of the measured confounder history.

Trivial example

Trivial example: choose $f^g(a_t|\bar{a}_{t-1}, \bar{l}_t) = f^{obs}(a_t|\bar{a}_{t-1}, \bar{l}_t)$ where $f^{obs}(a_t|\bar{a}_{t-1}, \bar{l}_t)$ is the factual (observed) treatment distribution in the population

$$\sum_{\bar{a}_K, \bar{l}_K} \mathrm{E}[Y|\bar{A}_K = \bar{a}_K, \bar{L}_K = \bar{l}_K] \prod_{t=0}^{K} f^{obs}(a_t|\bar{a}_{t-1}, \bar{l}_t) \prod_{t=0}^{K} f(l_t|\bar{a}_{t-1}, \bar{l}_{t-1})$$

This would be the intervention treatment distribution if g were simply chosen as "do nothing" or "assign treatment as it was assigned in fact" (*natural course*).

- By laws of probability, can show that the g-formula for this special choice reduces to simply the factual mean of Y $\mathrm{E}[Y]$
- For binary treatment, $f^{obs}(a_t = 1|\bar{a}_{t-1}, \bar{l}_t)$ is the so-called *propensity score*

Back to positivity

The positvity assumption is needed to ensure the g-formula is well-defined

$$\sum_{\bar{a}_K,\bar{l}_K} \mathrm{E}[Y|\overline{A}_K = \bar{a}_K, \overline{L}_K = \bar{l}_K] \prod_{t=0}^{K} f^g(a_t|\bar{a}_{t-1},\bar{l}_t) \prod_{t=0}^{K} f(l_t|\bar{a}_{t-1},\bar{l}_{t-1})$$

Requires that there are no histories within the sum, weighted positively by $f^g(a_t|\bar{a}_{t-1},\bar{l}_t)$, that are impossible to see in the factual world – that is, under $f^{obs}(a_t|\bar{a}_{t-1},\bar{l}_t)$. This avoids conditioning on empty strata.

Example: "Everyone exercise exactly 30 minutes everyday"

Deterministic interventions on continuous treatments may be particularly subject to positivity violations.

$$\sum_{\bar{l}_K} \mathrm{E}[Y|\overline{A}_K = \overline{30}_K, \overline{L}_K = \bar{l}_K] \prod_{t=0}^{K} f(l_t|\overline{30}_{t-1}, \bar{l}_{t-1})$$

If the number of minutes per day in the real world varies widely, positivity violations are inevitable for any deterministic g; e.g. there may be no individuals exercising exactly 30 minutes every day for 5 years.

- Can mitigate this with stochastic rules
- One particularly convenient choice is a "representative intervention"

Representative intervention

Suppose we select g such that
$f^{g}(a_t|\overline{a}_{t-1},\overline{l}_{t-1}) = f^{obs}(a_t|R_t = 1,\overline{a}_{t-1},\overline{l}_{t-1})$,

- where R_t is an indicator that treatment A_t is in a pre-specified range at t in the factual world.
- E.g. R_t indicator that individual exercised "at least 30 minutes"

This would be the form of the intervention treatment distribution if we chose the rule g at each t as:

- At each time t, for an individual with covariate history $\overline{a}_{t-1},\overline{l}_{t-1}$, randomly assign them a value of minutes of exercise on that day from the factual (observed) treatment distribution among those with the same history as that individual AND who actually exercised at least 30 minutes (or more generally within the pre-specified range).

This is a mouthful!

Representative intervention

This is not at all as natural a rule as “everyone exercise exactly 30 minutes everyday”. But it is a special case of a rule that sounds natural “everyone exercise at least 30 minutes everyday”. Further, it is less subject to positivity violations.

- But there are an infinite number of ways we could have more precisely defined a rule “everyone exercise at least 30 minutes everyday” and is less subject to positivity violations.
- We choose this one because it leads to particularly tractable statistical algorithms (Part III).

Questions?

Break

Part III: Statistics – estimating the g-formula and associated contrasts in a dataset

g-methods and the special case of inverse probability weighting

g-methods

There are number of statistical methods we can use for estimating contrasts in the g-formula indexed by different choices of treatment rule g.

- As a whole these are sometimes referred to as *g-methods*

Keeping everything thus far fixed (the question, the identifying assumptions, and, in turn, the g-formula), there are still different g-methods we can consider. They will differ by their "dart board properties" and also in their computational complexity.

parametric g-formula/g-computation

The parametric g-formula (g-computation) has been around the longest (Robins, 1986). This method assumes that the terms of g-formula can be correctly characterized by parsimonious parametric models.

parametric g-formula/g-computation

Algorithm:

- A model for $\mathrm{E}[Y|\overline{A}_K, \overline{L}_K]$. (outcome regression model)
- Models for joint distribution of L_t given past treatment and confounders for all t (can be a lot of models).
- And any models needed for the intervention treatment distribution (not needed for deterministic rules)
- Fit the models, and approximate the sum over all levels of confounders by simulating "lots" of treatment and confounder histories consistent with g using the model parameters.
  - at each t, simulate confounders from models then set treatment according to g.
- Use the outcome regression to estimate the outcome mean conditional on each simulated history and then average these estimates.

parametric g-formula/g-computation

- Advantages: computationally very easy to adapt to any rule, the only step that changes in how you set treatment in simulation. Also relies on familiar parametric modeling approaches so not a huge learning curve. IF the statistical assumptions it requires are correct, the method has low variability (tight scatter around the center of the dart board).
    - R package now available: *gformula*, documentation published in McGrath et al. *Patterns* (2020)
- Disadvantage: these statistical assumptions are generally extremely strong. When they are wrong, you may get tight scatter very far from the center.

Inverse probability weighting

Alternative approach comes from the fact that g-formula indexed by rule g has a weighted representation:

$$\mathrm{E}\left(Y \frac{\prod_{t=0}^{K} f^{g}(A_t|\overline{L}_t, \overline{A}_{t-1})}{\prod_{t=0}^{K} f^{obs}(A_t|\overline{L}_t, \overline{A}_{t-1})} \right)$$

Weight is 0 for anyone with treatment inconsistent with g at any follow-up time. Otherwise, it is the probability of receiving the treatment that person received given their measured past under g divided by this same probability but under no intervention. This representation motives "inverse probability weighted" estimators.

Weight denominator

Weight denominator depends on $f^{obs}(A_t|\overline{L}_t, \overline{A}_{t-1})$ for all t. For binary treatment, this is fully defined by the so-called *propensity score*:

$$\Pr[A_t = 1|\overline{L}_t, \overline{A}_{t-1}]$$

When many follow-up times and high-dimensional L_t, can estimate this using a model, e.g. a pooled over time logistic regression model. Unbiasedness/consistency depends on this being correctly specified.

Generic IPW algorithm

- Make copies of the data for each choice of g
- In each g specific copy, calculate the weight $\frac{\prod_{t=0}^{K} f^{g}(A_t|\overline{L}_t,\overline{A}_{t-1})}{\prod_{t=0}^{K} f^{obs}(A_t|\overline{L}_t,\overline{A}_{t-1})}$ for each individual based on model for denominator plugging in that person's data
- In each g specific copy, compute a weighted version of the sample mean of Y using these weights
- Take difference (or ratio) as causal effect estimate

Boostraps for 95% Cis

Inverse probability weighting for deterministic static rules

For special case of deterministic static rules reduces to

$$\mathrm{E}\left(Y\frac{\prod_{t=0}^{K} I(A_t = a_t^*)}{\prod_{t=0}^{K} f^{obs}(A_t|\overline{L}_t, \overline{A}_{t-1})}\right)$$

Weight is 0 for anyone with treatment inconsistent with g at any follow-up time. Otherwise, it is inverse of the probability of receiving the treatment that person received given their measured past (numerator is 1).

Positivity violations

This approach will not perform well for continuous treatments for a few reasons. First, you would lose most of the data! Nearly everyone will get a zero weight for a static rule.

$$\mathrm{E}\left(Y\frac{\prod_{t=0}^{K} I(A_t = a_t^*)}{\prod_{t=0}^{K} f(A_t|\overline{L}_t, \overline{A}_{t-1})}\right)$$

One way to mitigate this is by assuming a so-called *marginal structural model*.

Marginal Structural Models (MSMs) for static rules

An MSM for static rules assumes that we can write the g-formula indexed by any static rule $\overline{a} = (a_0, \ldots, a_K)$ as a function of $\overline{a}$.

Examples of MSMs for static rules

One example:

$$\mathrm{E}\left(Y\frac{\prod_{t=0}^{K} I(A_t=a_t)}{\prod_{t=0}^{K} f(A_t|\overline{L}_t,\overline{A}_{t-1})}\right)=\beta_0+\beta_1 cumavg(\overline{a}_K)$$

Another example:

$$\mathrm{E}\left(Y\frac{\prod_{t=0}^{K} I(A_t=a_t)}{\prod_{t=0}^{K} f(A_t|\overline{L}_t,\overline{A}_{t-1})}\right)=\beta_0+\beta_1 a_K+\beta_2 a_{K-1}+\ldots+\beta_{K+1}a_0$$

Under the assumption of an MSM, all I need to do is estimate the MSM coefficients β_0, β_1, etc. and I know the g-formula indexed by any static rule I can come up with! E.g. always exercise 40 minutes, 41 minutes, 19 minutes, 30 minutes or even static rules that require different minutes at each time – any combination that we see in the data will be used in the estimation.

Marginal Structural Models (MSMs) for static rules

$$\mathrm{E}\left(Y\frac{\prod_{t=0}^{K} I(A_t = a_t)}{\prod_{t=0}^{K} f(A_t|\overline{L}_t, \overline{A}_{t-1})}\right) = \beta_0 + \beta_1 cumavg(\overline{a}_K))$$

Benefit: No one will get a 0 weight now because EVERYONE in the study has data consistent with some static rule. Big benefit:

- No explicit need for copies and can use off the shelf software
- Weighted outcome regression with dependent variable Y, weights the IP weights, independent variable a function of the time-varying treatment $\overline{A}_K$ corresponding to chosen MSM functional form (e.g. cumavg, separate indicators for treatment status at each time, etc.)

Marginal Structural Models (MSMs) for static rules

$$\mathrm{E}\left(Y \frac{\prod_{t=0}^{K} I(A_t = a_t)}{\prod_{t=0}^{K} f(A_t | \overline{L}_t, \overline{A}_{t-1})} \right) = \beta_0 + \beta_1 cumavg(\overline{a}_K)$$

Downside: the MSM assumption is usually a pretty arbitrary parametric assumption and the more wrong it is, the more bias in your effect estimates.

- Beyond this, we problem of greater susceptibility to positivity violations with these types of rules.
- Another way to see this is by the structure of the weight – 1 over zero (or with "near positivity violations" something very close to zero)

Representative interventions

As we said before, stochastic interventions mitigate positivity violations

- Representative interventions “convenient” choice for questions about continuous treatments
- Ultimately allows a *coarsened* version of real exposure in place of exposure in algorithm.
- Details...

The g-formula indexed by a representative intervention

We get this by replacing $f^g(a_t|\overline{a}_{t-1}, \overline{l}_{t-1})$ in

$$\sum_{\overline{a}_K, \overline{l}_K} \mathrm{E}[Y|\overline{A}_K = \overline{a}_K, \overline{L}_K = \overline{l}_K] \prod_{t=0}^{K} f^g(a_t|\overline{a}_{t-1}, \overline{l}_{t-1}) \prod_{t=0}^{K} f(l_t|\overline{a}_{t-1}, \overline{l}_{t-1})$$

with $f^{obs}(a_t|R_t = 1, \overline{a}_{t-1}, \overline{l}_{t-1})$ which we can write as:

$$\sum_{\overline{a}_K, \overline{l}_K} \mathrm{E}[Y|\overline{A}_K = \overline{a}_K, \overline{L}_K = \overline{l}_K] \prod_{t=0}^{K} f^{obs}(a_t|R_t = r_t, \overline{a}_{t-1}, \overline{l}_{t-1})$$
$$\prod_{t=0}^{K} f(l_t|\overline{a}_{t-1}, \overline{l}_{t-1})$$

when we select $r_t = 1$ at all t.

Weighted representation of the g-formula indexed by a representative intervention

It turns out that the IP weighted representation of this special case of the g-formula is

$$\mathrm{E}\left(Y \frac{\prod_{t=0}^{K} I(R_t = r_t)}{\prod_{t=0}^{K} f(R_t | \overline{L}_t, \overline{A}_{t-1})} \right)$$

Does this look familiar? Recall the form for a static rule

$$\mathrm{E}\left(Y \frac{\prod_{t=0}^{K} I(A_t = a_t)}{\prod_{t=0}^{K} f(A_t | \overline{L}_t, \overline{A}_{t-1})} \right)$$

It turns out that the IP weighted representation of the g-formula for this special stochastic rule LOOKS just like that for a static rule! This means we can rely on the simplicity of IP weighted estimators for static rules and corresponding MSMs – just replace A_t with R_t!

Marginal Structural Models (MSMs) for representative interventions

A marginal structural model in this case can be written as a function of $\bar{r}_t$. E.g.

$$\mathrm{E}\left(Y \frac{\prod_{t=0}^{K} I(R_t = r_t)}{\prod_{t=0}^{K} f(R_t|\overline{L}_t, \overline{A}_{t-1})} \right) = \beta_0 + \beta_1 cumavg(\bar{r}_K)$$

Can get the coefficients similarly by a weighted linear regression! Dependent variable Y, independent variables a function of time-varying indicators of being in the pre-specified treatment range, weights the IP weights, denominator can just be fit with a propensity score like model.

- Can get mean under set e.g. “always exercise at least 30 minutes” by predicting the mean from this model plugging in all $r_t = 1$.
- Might compare to $r_t = 0$ at all times (“never exercise at least 30 minutes”.)

Near positivity violations

Even with representative interventions, we may run into so-called "near positivity violations" –

- there are individuals in the data who have data consistent with g (so weight not zero) but given their confounders they are very unusual (few people like them who have data consistent with g)

The weight for these folks will be 1/something really, really small. In these settings, IPW has pretty terrible "dart properties".

- We can usually see this by looking at weight distributions.

Stabilized weights can help.

Unstabilized versus Stabilized weights and conditional MSMs

Thus far we have considered the unstabilized weights for each individual

$$\frac{\prod_{t=0}^{K} I(R_t = r_t)}{\prod_{t=0}^{K} f(R_t|\overline{L}_t, \overline{A}_{t-1})}$$

. We have the option to instead use weights

$$\frac{\prod_{t=0}^{K} f(R_t|V, \overline{R}_{t-1}) I(R_t = r_t)}{\prod_{t=0}^{K} f(R_t|\overline{L}_t, \overline{A}_{t-1})}$$

where V can include any (or all) components in L_0. This can help mitigate extreme weights.

- However, to validly use stabilized weights with V in the numerator, the MSM must be conditioned on the same V

A tradeoff

The more we include in V, the more in principle we may stabilize the weights (bring numerator closer to denominator).

- But to use stabilize weights depending on V, the MSM must be conditioned on V and thus makes stronger assumptions

$$\mathrm{E}\left(Y\frac{\prod_{t=0}^{K} I(A_t=a_t)}{\prod_{t=0}^{K} f(A_t|\overline{L}_t,\overline{A}_{t-1})}|V\right)=\beta_0+\beta_1 a_K+\beta_2 a_{K-1}+\ldots+\beta_{K+1}a_0+\gamma^T V$$

Can get back overall effects (marginal over V) at the end if we want by "averaging out V".

Stabilized weights: Why do they work?

Technical reason: IPW is a solution to an estimating equation that must have mean zero to be consistent. In the case of a static MSM conditioned on V, adding a function of time-varying treatment and V to weight numerator number boils down to multiplying the estimating equation by a constant that has no effect on its mean.

- Same idea for representative interventions where, in the algorithm, R_t (a “coarsened” version of treatment) acts like the actual treatment.

Part IV: When counterfactual contrasts are not causal effects

The problem of conditioning on post-treatment variables

What's wrong with standard/familiar regression methods?

Familiar regression is an option when we are interested in causal effects of *time-fixed static deterministic treatment rules* and we want *conditional effects*.

- If we wanted to know the causal effect of "treat" versus "do not treat" among any level of the measured confounders L_0, we could rely on an outcome regression model for $\mathrm{E}[Y|A=a, L_0=l_0]$. If this model was correctly specified (and our causal reasoning to get here was right, then the coefficient on a in this model is a consistent/unbiased estimate of that effect $\mathrm{E}[Y^{a=1}|L_0=l_0] - \mathrm{E}[Y^{a=0}|L_0=l_0]$.
- That model assumption gets very strong when there is a lot in L_0
- Problem with standard regression methods gets even worse when we have time-varying treatments.

Return to our idealized trial

Interest is in “always take medication” versus “never take”. Consider simple case of 2 timepoints and we conduct the idealized trial to answer this so that we can identify our effect of interest which is simply by

$$E[Y^{\bar{a}=\bar{1}}] - E[Y^{\bar{a}=\bar{0}}]$$

with

$$E[Y|Z=1] - E[Y|Z=0]$$

. Now suppose that the investigator of this trial says, I'm interested, not in the marginal effect but in a conditional effect, conditional on some L_1 that happens after A_0 and can be affected by A_0. They then try to estimate this by estimating

$$E[Y|Z=1, L_1] - E[Y|Z=0, L_1]$$

.

The problem of conditioning on L_1

The investigators logic is problematic because the stated interest in "the same causal effect but conditional on L_1" is not comprehensible because it is not a causal effect.

- In the $Z = 1$ arm, where everyone gets $A_0 = 1$ L_1 means $L_1^{a_0=1}$
- In the $Z = 0$ arm, where everyone gets $A_0 = 0$ L_1 means $L_1^{a_0=0}$

When counterfactual contrasts are not causal effects

In turn, even in this idealized study, an unbiased estimate of

$$E[Y|Z=1,L_1]-E[Y|Z=0,L_1]$$

is actually an estimate of

$$E[Y^{\bar{a}=\bar{1}}|L_1^{a_0=1}=l_1]-E[Y^{\bar{a}=\bar{0}}|L_1^{a_0=0}=l_1]$$

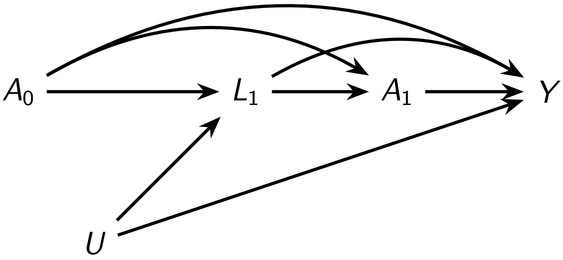
This is not a causal effect – it compares outcomes under different treatments in different individuals

- These are only the same individuals when A_0 does not affect L_1 which counters our original subject matter knowledge.

When counterfactual contrasts are not causal effects

It may be the case that we could clarify with this investigator what they really want to know – help them to articulate an actual causal effect that somehow involves L_1

- For example, maybe they really want some direct effect of not involving paths that includes L_1
- Many ways to define this
- To identify a direct effect, we will require assumptions that are not guaranteed even in this idealized study
- E.g. we cannot identify any such direct effect if L_1 and Y share unmeasured common causes



Conditional on a particular level of L_1, A_0 and U are associated – if both treatment A_0, say, improves your level of L_1, and so does U (maybe a protective gene) then if you take two people, one from the treated arm and one from the no treated arm, the no treated arm more likely to have the gene (the arms are no longer balanced on background causes of outcome).

- Phenomenon sometimes called “collider bias” in literature

Death is a typical L_1

This argument is precisely why, in a study of say effects of prenatal or preconception treatments on offspring outcomes, an analysis that conditions on livebirths and compares outcomes across treatments, with no adjustment for common causes is biased for causal effects, even in a perfectly executed trial!

- Imperative to be explicit about what we want to know in these settings.

Death is a typical L_1

- It can be very challenging to even articulate questions in this setting because often what we really want are direct effects – not through death
- Historical notions of direct effects require ill-defined intervention on death or restricted to unidentifiable "always survivors"
- New definitions overcome these problems – the *separable effects* – which require conceptualizing modified treatments that remove harmful mechanism of the current study treatment.
- Always rely on stronger assumptions than the ones we've reviewed today – which all restrict to *total effects* capturing all mechanisms by which implementing a g_1 versus a g_2 may affect outcomes .

Refer to readings for detailed debate and discussion: Chiu et al, Snowden et al, Young & Stensrud

Important issues/topics I did not adequately discuss (quick summary/thoughts)

Censoring events

- Any event that creates missingness in (possibly counterfactual) outcomes *of interest*
- One approach, implicitly define all questions in terms of additional intervention "eliminate censoring"
- Censoring indicator at each time is then another treatment
- If we impose additional exchangeability, consistency, positivity assumptions for this additional "static rule", comparable IP censoring weights fall out (Yu-Han will illustrate next!)

Doubly robust/"State of the art" estimators

Another representation of the g-formula as a series of iterated conditional expectations (means) motivate doubly robust methods – weaken statistical assumptions in that only requires correct specification of one of two sets of parametric models.

- Models for weight denominator
- Models for iterated conditional means

Unlike the "singly robust" methods above, adaptations of these approaches can further weaken statistical assumptions so that flexible machine learning methods can be used in place of parametric models.

- Aris et al. in included readings includes simple application.

DR/State of the art methods

Selected software resources for g-formula contrasts indexed by some of the types of effects we have reviewed.

- *tmle* R package – point treatments, Gruber
- *ltmle* R package – time-varying treatments, deterministic interventions, Petersen et al.
- *lmtp* R package – time-varying continuous treatments, stochastic interventions, Díaz-Muñoz
- E. Kennedy – https://github.com/ehkennedy/npcausal, stochastic interventions for binary treatments

Wide measurement intervals

- Often we cannot measure the exposure of interest as frequently as we would ideally intervene to answer our question
- Particularly true of non-medical exposures which are often collected via questionnaire
- E.g. we may only measure changes in exercise or diet at the end of a yearly interval based on self-report when our question is really about daily interventions
- Of course more assumptions are needed in this case
- Can think through identification in a few ways
- One is to think of this yearly measure as another proxy for the daily exposure in that period and make a type of coarsening assumption (see appendix of Aris et al).

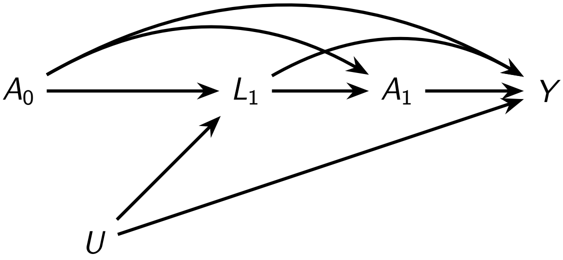
Questions?

Using SWIGs to evaluate exchangeability: some examples

For extremely committed attendees, see Richardson and Robins (2013) for details: https://csss.uw.edu/research/working-papers/single-world-intervention-graphs-swigs-unification-counterfactual-and

- In particular, Theorem 31, page 67, and Corollary 34, page 71

How to evaluate exchangeability.

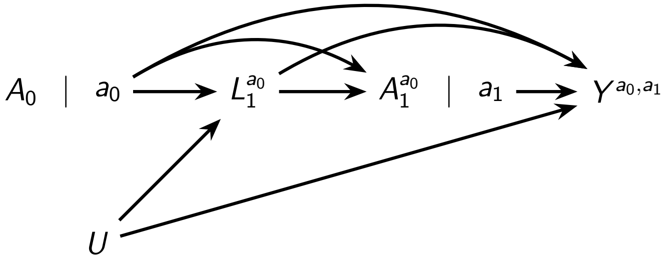


Recall exchangeability is independence:

$$Y^g \coprod A_t | \overline{L}_t, \overline{A}_{t-1} = \overline{a}^g_{t-1}$$

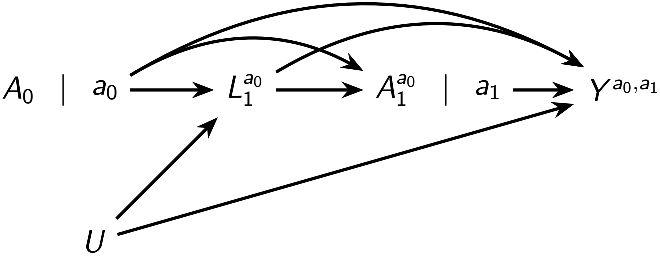
Specific to g, not clear how this links to this graph? No "g" on it, no counterfactuals. Evaluating this historically required implicit transformations that led to reasoning errors. Single World Intervention Graphs solve this .

SWIG – static rule "Set A_t to constant a_t at all times", denote $\overline{a}$, special case of g



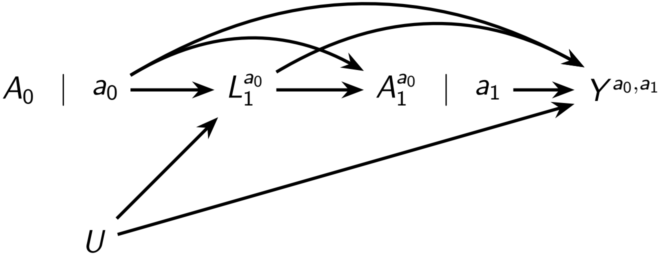
SWIG under g transforms the causal DAG in a particular way: split treatment nodes into "natural treatment value" and "intervention value". All nodes going into treatment on DAG go into natural value on SWIG. All nodes out on DAG go out of intervention value. Natural treatment value at t = treatment value at t had intervention been conducted through $t-1$ but none at t.

SWIG – static rule "Set A_t to constant a_t at all times", denote $\overline{a}$, special case of g



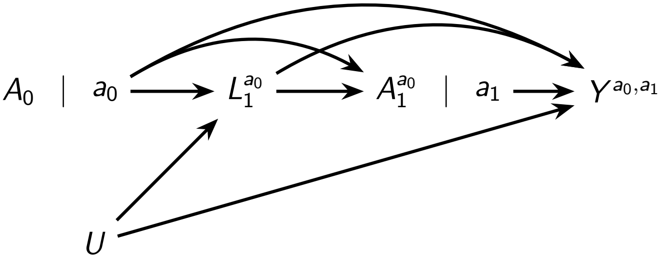
A_0 is both factual (observed) and natural value under g because pre-intervention. A_1 (factual) is not necessarily the same as natural value $A_1^{a_0}$ but can link through consistency for people with history $A_0 = a_0$ in the real world.

SWIG – static rule “Set A_t to constant a_t at all times”, denote $\bar{a}$, special case of g



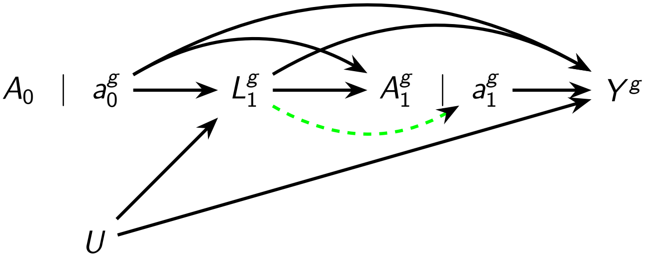
If dynamic g add dashed arrows into intervention value from L’s.

SWIG – static rule “Set A_t to constant a_t at all times”, denote $\overline{a}$, special case of g



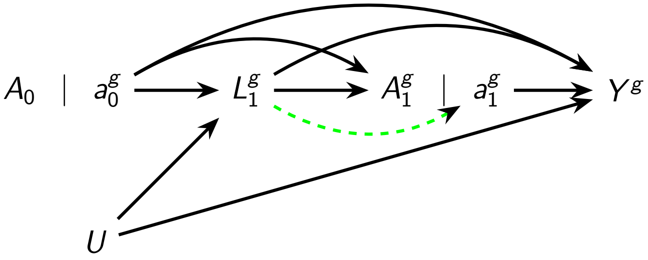
Evaluate 1) Any “open backdoor paths” between A_0 and Y^{a_0,a_1}? 2) Any “open backdoor paths” between $A_1^{a_0}$ and Y^{a_0,a_1} conditional on $L_1^{a_0}$, A_0, a_0. Invoking consistency and b/c a_0 a constant, this implies exchangeability as we stated it.

SWIG – dynamic rule g example, "Treat at time 0. If no contraindication developed then treat at time 1, o.w. don't treat at time 1"



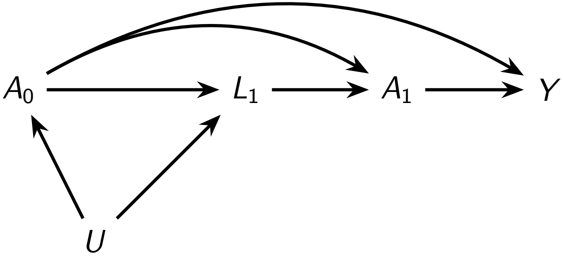
Differs from static example visually in how we label counterfactuals but also green arrow indicates that the intervention level of treatment at time 1 depends on the level of L_1 (under g).

SWIG – dynamic rule g example, "Treat at time 0. If no contraindication developed then treat at time 1, o.w. don't treat at time 1"



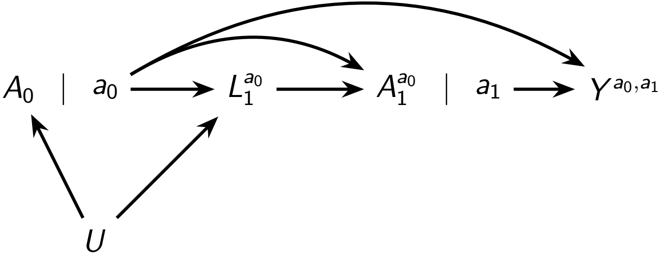
Evaluate 1) Any "open backdoor paths" between A_0 and Y^g? 2) Any "open backdoor paths" between A_1^g and Y^g conditional on L_1^g, A_0, a_0^g. Invoking consistency and b/c a_0^g constant, this evaluates exchangeability as we stated it.

Example where green arrow matters



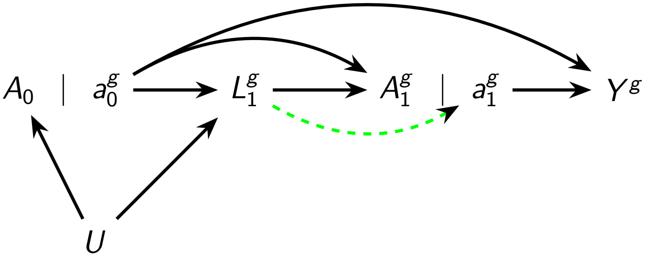
This causal DAG differs from other one in that 1) switched arrow into Y to be into A_0 instead and 2) removed the arrow from L_1 to Y.

Example where green arrow matters



Evaluate 1) Any "open backdoor paths" between A_0 and Y^{a_0,a_1}? 2) Any "open backdoor paths" between $A_1^{a_0}$ and Y^{a_0,a_1} conditional on $L_1^{a_0}$, A_0 and a_0.

Example where green arrow matters



Evaluate 1) Any “open backdoor paths” between A_0 and Y^g? 2) Any “open backdoor paths” between A_1^g and Y^g conditional on L_1^g, A_0 and a_0^g.