# CAUSAL INFERENCE IN POINT-TREATMENT AND LONGITUDINAL STUDIES

## LECTURE I:

## INTRODUCTORY STATEMENTS AND OVERVIEW OF COURSE

# POINT TREATMENT

Causal inference distinguishes between a study with treatment being time-independent and longitudinal studies with time-dependent treatment.

One is often concerned with estimation of a *causal effect* (a parameter with a causal interpretation) of a variable which can be manipulated (Exposure or Treatment) on an outcome of interest, possibly *adjusted* for other variables.

**Example:** Estimate the (adjusted) causal effect of being a current cigarette smoker on the level of forced expiratory volume in one second (FEV1) in a cohort of 2713 adult white male former and current cigarette smokers from cross-sectional data collected in the Harvard Six Cities Study (Dockery et al., 1988).

See table that includes variables on past smoking history, past respiratory symptoms, age, height and coexistent heart disease.

# A CAUSAL MODEL

Data Generating Experiment: Randomly draw subject from population, measure baseline covariates $W$, assign/measure treatment/exposure variable $A$ and measure the outcome of interest. The data on a randomly selected subject is $(Y, A, W)$.

Let $Y_a$ be the random variable $Y$ one would have observed, if, possibly contrary to the fact, one would have "assigned" $A = a$. One refers to $Y_a$ as a *counterfactual variable*. The *counterfactual distribution/treatment specific distribution* of $Y_a$ is the distribution one would observe in the hypothetical experiment in which we set $A = a$ for each subject in the population we draw from.

**Linking counterfactuals to the observed data:** Each subject has an underlying vector of counterfactuals $(Y_{a,i}, a \in \mathcal{A})$. If subject $i$ has been assigned exposure/treatment $A_i$ in the actual study, then his/her observed $Y_i$ equals $Y_{A_i}$. The other $Y_{a,i}$, $a \neq A_i$, are all missing.

Thus one observes $(A_i, Y_i = Y_{A_i}, W_i)$ on each subject.

A **causal model** involves modelling of the effect of $a$ on $Y_a$, possibly adjusted for $V \subset W$.

An example of a causal model: $E(Y_a \mid V) = \beta_0 + \beta_1 a + \beta_2 V$. In this *causal linear regression model* $\vec{\beta}$ is a **causal parameter**.

# ASSOCIATION VERSUS CAUSALITY

**Regression model for observed data:**

$$E(Y \mid A) = \alpha_0 + \alpha_1 A.$$

**Causal regression model:** For all treatment outcomes $a$

$$E(Y_a) = \beta_0 + \beta_1 a.$$

If $\vec{\alpha} = \vec{\beta}$, i.e. if

$$E(Y \mid A = a) = E(Y_a),$$

then the regression parameters $\vec{\beta}$ are causal parameters and we say that there is no confounding.

If $E(Y \mid A = a) \neq E(Y_a)$, then we say that the effect of $A$ on $Y$ is *confounded*.

**Confounding in terms of propensity score:** We define

$$P(A = a \mid \text{subjects characteristics})$$

as the *propensity score*. Formally, the subjects characteristics are defined by $\{Y_a : a \in \mathcal{A}\}$ and the measured covariates.

In words, it equals the probability on a particular treatment, given the subject.

In a study where one collects $(Y, A)$ on each subject, but no additional covariates, we say that $A$ is randomized if

$$P(A = a \mid \text{subjects characteristics}) = P(A = a).$$

In a study where one collects $(Y, A, W)$ on each subject we say that $A$ is randomized if

$$P(A = a \mid \text{subjects characteristics}) = P(A = a \mid W).$$

In words: the treatment variable is randomized if the probability on a particular treatment outcome is a function of the observed covariates only.

One also refers to this assumption as the *assumption of no unmeasured confounders*.

If $A$ is randomized in a study collecting data $(Y, A)$, then

$$E(Y \mid A = a) = E(Y_a).$$

If $A$ is randomized in a study collecting data $(Y, A, W)$, then

$$E(Y \mid A = a, W) = E(Y_a \mid W).$$

**Classic example of confounding:** "Carrying matches" is associated with lung cancer, but "carrying matches" does not cause lung cancer.

# OBSERVATIONAL VERSUS RANDOMIZED.

In a *randomized study* (e.g. clinical trial) the assignment of treatment is under control of the experimenter. In this case the propensity score is known.

In an *observational study* the propensity score is unknown, but one can still hope/arrange that the assumption of no unmeasured confounder holds by collecting as many potential confounders as possible.

# EXAMPLE

Consider a study involving pregnant women and let the outcome $Y$ of interest be the indicator of a birth defect.

**Data:** On $n$ subjects we observe $Y$ and several variables of interest such as the level $A$ of alcohol consumption and smoking.

**Question**: Does smoking/alcohol consumption have a *causal effect* on the presence of a birth defect? In other words, if we would force each women in the population to stop smoking and drinking during pregnancy, would that decrease the number of birth defects?

**Linear regression approach:** Assume

$$Y = \alpha_0 + \alpha_1 A + \text{error}.$$

Estimate $\alpha_1$ with linear regression of $Y$ on $A$.

**Confounding:** Large percentage of woman who smoke and drink have stressful jobs and bad eating habits. Thus even when there is no causal effect of smoking/drinking one might find that $\widehat{\alpha} > 0$.

**Adding confounders to the linear regression model?** This is not solving the question!

**Key to solution:** Use causal linear regression model: For each smoking/drinking level $a$ let $Y_a$ be a random variable whose distribution equals the population distribution of $Y$ if each subject would smoke/drink at level $a$.
Model dependence of $Y_a$ on $a$: For example, assume $Y_a = \beta_0 + \beta_1 a +$error and estimate $\beta_1$.

# EXAMPLE

**Breast Cancer Data** A clinic in Germany collected data on 225 women with breast cancer. At the time of detection, the tumor was surgically removed and variables were recorded that are believed to reflect the progression and severity of disease (for example, tumor size, tumor type and the number of lymph nodes involved). After surgery, each woman either received chemotherapy or not. The time until tumor recurrence is the outcome of interest and it is subject to right-censoring.

**Question:** Does chemotherapy have a causal effect on time till tumor recurrence? Would the time till recurrence distribution improve if each woman would receive chemotherapy?
**Association method:** Compare survival (Kaplan-Meier) estimate in treatment group with sur-

vival estimate in non-treatment group.

**Confounding:** Women with a poorer prognosis were more likely to receive aggressive treatment, i.e. chemotherapy.

**Causal method:** Estimate treatment specific population distributions.

# TOPICS RELEVANT FOR THIS COURSE

- Graphical conditions for identifying a causal effect. Confounding defined by graphical criteria.

- Nonparametric structural equation model for a graphical model.

- Direct and indirect effects (Robins and coworkers)

- Non compliance in randomized studies (Robins and coworkers).

- Marginal Structural Models: Estimation and Inference (Robins).

Papers:

1) “Causal diagrams in epidemiologic research” by Greenland, Pearl and Robins (1998).

2) “Causal diagrams in empirical research” by Pearl (1995) with discussions.

3) “Why there is no statistical test for confounding, why many think there is, and why they are almost right” (Pearl, 98).

4) “Statistics, Causality and Graphs” (Pearl, 97).

5) “Marginal Structural Models” (Robins, 98)

6) “Estimating Exposure Effects by modelling the expectation of exposure conditional on confounders” (Robins, Mark, 92).

## Longitudinal Studies.

In a longitudinal study one collects data on a subject over time. Let $A(\cdot)$ be a treatment process, where $A(k)$ denotes the treatment the subject receives at time $k \in \{1, 2, 3, 4, \ldots\}$. Let $Y(\cdot)$ be an outcome process where $Y(k)$ denotes the outcome measured between time $A(k-1)$ and $A(k)$, preceding $A(k)$. Let $L(\cdot)$ be a covariate process, where $L(k)$ represents the covariates measured between time $A(k-1)$ and $A(k)$, preceding $A(k)$.

The data generation process can be thought of as a sequence of experiments over time, where the experiment at time $k$ is conditional on the observed past. Treatment is now *seqeuntially randomized* if the treatment assignment $A(k)$ in experiment $k$, conditional on the past, is randomized (defined as in the point exposure study). In other words, the treatment assignment $A(k)$ is only based on the data available at that point in time: i.e. $A(1), \ldots, A(k-1)$, $L(1), \ldots, L(k)$, $Y(1), \ldots, Y(k)$.

Causal Inference in longitudinal studies is very delicate if there exist time-dependent covariates which predict future treatment (i.e. are a potential confounder) and are on the causal pathway from treatment to the outcome: Make a picture: 1) treatment(1) effects covariate(2), 2) covariate(2) effects treatment(2) and future outcome etc.

Give example (treatment, cholesterol and heart disease) of point-exposure study where one uses the $G$-computation formula adjusting for a variable on the causal pathway from $A$ to $Y$, showing that the $G$-computation formula gives a useless answer.

**Example I**: Consider a study of the effect of post-menopausal oestrogen on cardiac mortality in which one collects as time-dependent covariate the cholesterol level.

Cholesterol level predicts cardiac mortality.

Cholesterol level also predicts future treatment since physicians withdraw women from oestrogens at the time they develop an elevated cholesterol level.

Being on oestrogens might effect the future cholesterol level.

**Example II**: Consider an observational study of the efficacy of breast cancer screening (treatment/exposure) on mortality in which one collects also the time-dependent covariate "operative removal".

Operative removal predicts mortality.

After operative removal the screening (treatment) stops.

"Operative removal" is on the causal pathway from “Being screened” to death.

**Example III:** Consider an observational study of the effect of AZT-treatment on times to AIDS in HIV-infected subjects in which CD4-count is a measured time-dependent covariate.
CD4 predicts death and treatment and is on the causal pathway from AZT to time till AIDS.

# QUESTIONS OF INTEREST.

- The difference between (parameters of) the treatment specific outcome distributions corresponding with "never treat" and "always treat" , possibly adjusted for baseline covariates.

- Estimation of the treatment specific outcome distributions corresponding with a given set of possible treatment stategies, possibly dynamic treatment stategies, possibly adjusted for baseline covariates.

- Optimal treatment strategy.

- Given a subject made it up till point $t$ and given its covariate and treatment history up till point $t$, what is the difference between the treatment specific outcome distributions corresponding with "treating at point $t$ and never after" and "not treating at point $t$ and never after".

## TOPICS ADDRESSED IN THIS COURSE

- Marginal Structural Models.

- Structural Nested Models.

**Papers:**

1) The control of confounding by intermediate variables (Robins, 89).

2) Estimation of effects of sequential treatments by reparametrizing directed acyclic graphs (Robins, Wasserman, 98). 3) Marginal structural models and causal inference in epidemiology (Robins, 1999).

4) Structural nested failure time models (Robins, 97). 5) Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (Robins, Hu,

1993).

6) Estimation of the time-dependent accelerated failure time model in the presence of confounding factors.

7) G-estimation of causal effects: Isolated Systolic Hypertension and cardiovascular death in the Framingham study (Witteman et al. 1998).

8) Adjusting for differential rates of prophylaxis therapy for PCP in high versus low-dose AZT treatment arms in an AIDS randomized trial (Robins, Greenland, 1993).

9) G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients (Robins et al., 1992).

10) Correcting for non-compliance in randomized trials using rank preserving structural nested failure time models (Robins, 1991).

11) Correction for non-compliance in equiva-

lence trials (Robins, 1998).

# PART II: CAUSAL GRAPHS

# CAUSAL GRAPH RESEARCH

Pearl (1995) developes a formal theory for evaluating and identifying causal effects of single treatment variables using the language of causal graphs.

Robins (many papers) provides an actual formula for the counterfactual distributions in terms of the observed data distribution in longitudinal studies under the assumption of sequential randomization. This formula is called the $G$-computation formula which is very simple in the single treatment case.

The following lectures are concerned with showing how diagrams can serve as a visual yet logically rigorous aid for 1) summarizing assumptions about a problem, 2) identifying variables that must be measured and controlled to obtain unconfounded effect estimates.

# GRAPH TERMINOLOGY

Consider the graph in Figure 1. In this example, $A$ is air-pollution level, $B$ is sex (boy or girl), $C$ is bronchial activity, $E$ is antihistamine treatment, $D$ is astma.

ARC, EDGE: line or arrow connecting two variables.

ADJACENT: $A$ and $C$ are adjacent.

Single headed arrows represent direct links from causes to effects.

NODES.

PATH is any unbroken route traced out along or against arrows or lines connecting adjacent nodes: e.g. E-C-D is a path.

DIRECTED PATH/ CAUSAL PATH

node INTERCEPTS the path.

$X$ is an ANCESTOR or CAUSE of $Y$ if there is a directed path from $X$ to $Y$.

Then $Y$ is a DESCENDANT of $X$ or AFFECTED by $X$.
X PARENT of Y.
Y CHILD of X, $X$ is DIRECTLY AFFECTED by $Y$.

Unspecified common ancestors are denoted with $U$, with dashed arrows to the variables it affects.
DIRECTED GRAPH: all arcs between variables are arrows (single or double headed).
ACYCLIC GRAPH: no directed path forms a closed loop.
Abbreviation for directed acyclic graph: DAG.
A path that connects $X$ to $Y$ is a BACK DOOR PATH from $X$ to $Y$ if it has an arrowhead pointing to $X$. Figure 1: all path from $E$ to $D$ except the direct path are back door paths.

A path COLLIDES at a variable $X$ if the path enters and exits $X$ through arrowheads, in which case $X$ is called a collider on the path. A path is BLOCKED if it has one or more colliders, otherwise UNBLOCKED.
See figure 1: the back door path EACBD is blocked because it collides at $C$.
E-A-C-D is unblocked. CAUSE: $A$ is a cause of $C$.

**Definition:** A directed acyclic graph $G$ is a CAUSAL GRAPH if for each node $X_i$ with parents $(PA)_i$ we have $X_i = f_i((PA)_i, \epsilon_i)$ with $f_i$ being a deterministic function and $\epsilon_i, i = 1, \ldots, m$, are all independent, and $\epsilon_i$ is also independent of $(PA)_i$, $i = 1, \ldots, m$.

Let $A, Y$ be two nodes in the causal graph, where we have an arrow going from $A$ to $Y$. The counterfactual distribution of $Y_a$ is defined by
1) delete the equation corresponding with $X_i = A$.
2) Set $A = a$ in all the other equations.

Let $(L, U)$ represent all non-descendants of $A$. In a causal graph we have that $P(A = a \mid (Y_a, a \in \mathcal{A}), L, U) = P(A = a \mid L, U)$, i.e. $A$ is randomized w.r.t. observing the whole graph.

**DEFINITION OF CONFOUNDING:** In a causal DAG we say that the effect of $A$ on $Y$ is confounded if there is an unblocked back door path from $A$ to $Y$.

# STATISTICAL GRAPH

Let $X_1, \ldots, X_m$ be $m$ variables. Suppose $f(x_i \mid x_1, \ldots, x_{i-1}) = f(x_i \mid (pa)_i)$, where $(pa)_i$ is a subset of $(x_1, \ldots, x_{i-1})$. If we refer to $(x_1, \ldots, x_{i-1})$ as the ancestors of $x_i$, then this says that $X_i$ is independent of its ancestors, given its parents $(PA)_i$. In this case the density of $(X_1, \ldots, X_m)$ is given by:

$$p(X_1, \ldots, X_m) = \prod_{i=1}^{m} p(X_i \mid (PA)_i).$$

This likelihood of $(X_1, \ldots, X_m)$ corresponds with a STATISTICAL GRAPH defined by the nodes $X_1, \ldots, X_n$, where node $X_i$ has incoming arrows from $(PA)_i$.

**Remark:** A causal graph is also a statistical graph. A statistical graph is not necessarily a causal graph.

## d-SEPARATION IN STATISTICAL GRAPH

**d-SEPARATION:** Let $R$, $T$ and $S$ be three sets of nodes in the graph. We say that $R$ and $T$ are d-separated by $S$ if every unblocked path, including paths generated by adjustment for variables in $S$, from $T$ to $R$ is intercepted by a variable in $S$.

We can also say: $S$ blocks every path between $R$ and $T$.

In a statistical graph we have that $\vec{Z}_1$ is independent of $\vec{Z}_2$, given a third vector $\vec{Z}_3$ (all three vectors should be distinct) if $\vec{Z}_1$ and $\vec{Z}_2$ are d-SEPARATED by $\vec{Z}_3$.

The converse is not necessarily true: Figure 1 has a direct path and four back-door paths between $E$ and $D$. Each path transmits an association, but these associations might cancel one another out. However, this always involves perfect cancellations so that for all practical purposes one is allowed to read "A and B are d-separated by $C$" as "A and B are independent, given $C$".

One says that the joint distribution $p(X_1, \ldots, X_m)$ is **faithfull** to the statistical graph if we have that $\vec{Z}_1$ is independent of $\vec{Z}_2$, given a third vector $\vec{Z}_3$ IF AND ONLY IF $\vec{Z}_1$ and $\vec{Z}_2$ are d-SEPARATED by $\vec{Z}_3$.

See figure 1 and 3 for a graphical illustration for the following: Marginally $A$ and $B$ are not associated since $A$ and $B$ are $d$-separated, but $A$ and $B$ are associated within stata of $C$.

# SUFFICIENT SET OF ADJUSTMENT

Let $A, Y$ be two nodes in the statistical graph. Let $L$ be a set of other nodes in the graph, being **non-descendents** of $A$. Denote the remaining non-descendents of $A$ with $U$.

Let $b(y \mid a)$ be the $G$-computation formula (Robins):

$$b(y \mid a) = \int p(y \mid a, l, u) dP(l, u).$$

If $A$ is randomized for the data $(Y, A, L, U)$, i.e. $A$ is independent of $Y_a$, given $L, U$, for each $a$, then $b(y \mid a) = P(Y_a = y)$. This holds, in particular, if $G$ is a causal graph.

However, suppose $U$ is not observed. Then this $G$-computation formula is not useful because it cannot be estimated from data. Therefore it is of interest to understand under what conditions we have **that $L$ is a sufficient set of adjustment**: i.e.

$$b(y \mid a) = b^*(y \mid a) \equiv \int p(y \mid a, l) dP(l).$$

**Back door path condition:** We say that there is no back door path from $A$ to $Y$ if $Y$ is $d$-separated from $A$ in $G_{\underline{A}}$, where $G_{\underline{A}}$ is the graph obtained from $G$ by deleting all outgoing arrows from $A$.
Notation: $A \perp_d Y$.
We say that there is no back door path from $A$ to $Y$, controlling for $L$, if $Y$ and $A$ are d-separated by $L$ in $G_{\underline{A}}$.
Notation: $A \perp_d Y \mid L$.

**Theorem** If there is no back door path from $A$ to $Y$ controlling for $L$, then $L$ is sufficient for adjustment: i.e.

$$b(y \mid a) = b^*(y \mid a).$$

# ALTERNATIVE CRITERIA

**Theorem** $U$ can be split up in $U_1, U_2$ where $U_1 \perp_d A \mid L$ in $G$ (choose $U_1$ maximal set) and $U_2 \perp_d Y \mid (A, L, U_1)$ in $G$

$$\Longleftrightarrow$$

$Y \perp_d A \mid L$ in $G_{\overline{A}}$, i.e. there is no back door path from $A$ to $Y$ controlled for $L$.

So a statistical graph can be used to determine a sufficient set of variables $L$ to adjust for to compute $b(y \mid a)$. However, then we still wonder if $b(y \mid a) = P(Y_a = y)$? We know that this is true if $P(A = a \mid (Y_a, a \in \mathcal{A}), L) = P(A = a \mid L)$ (the randomization assumption holds). This assumption holds if the statistical graph happens to be a causal graph, but if it is not, then this is still an open question.

## TWO APPROACHES

Therefore we have the following two approaches for determing a correct formula for $P(Y_a = y)$ using graph theory:

**Statistical Graph:** Using the statistical graph, determine a sufficient set of variables $L$ to adjust for, i.e. such that there is not back door path from $A$ to $Y$ controlled for $L$. This guarantees that $b(y \mid a) = b^*(y \mid a)$.

Now, just assume/hope/reason that the randomization assumption holds $P(A = a \mid (Y_a, a \in \mathcal{A}), L) = P(A = a \mid L)$. Then the $G$-computation formula $b^*(y \mid a)$ only adjusting for $L$ equals $P(Y_a = y)$.

**Causal Graph:** Using a causal graph (thus needing a much stronger set of assumptions pertaining a causal graph), determine a sufficient set of variables $L$ to adjust for, i.e. such that there is not back door path from $A$ to $Y$ controlled for $L$. Then the $G$-computation formula $b^*(y \mid a)$ only adjusting for $L$ equals $P(Y_a = y)$.

Note that the statistical graph theory is applicable under fewer assumptions, but if one is able to assume a causal graph, then that guarantees selection of a sufficient set of variables $L$ to truly estimate $P(Y_a = y)$.

# STATISTICAL CRITERIA.

The graphical condition *"$U$ can be split up in $U_1, U_2$ where $U_1 \perp_d A \mid L$ in $G$ and $U_2 \perp_d Y \mid (A, L, U_1)$ in $G$"* for $b(y \mid a) = b^*(y \mid a)$ is a little stronger than needed since $b(y \mid a) = b^*(y \mid a)$ is only a statement in terms of distributions. The following theorem for determining if $b(y \mid a) = b^*(y \mid a)$ assumes only a purely statistical assumption.
**Theorem** (Statistical criteria) If $U$ can be split up in $U_1, U_2$ where $U_1$ is independent of $A$, given $L$ and $U_2$ is independent of $Y$, given $(A, L, U_1)$, then $b(y \mid a) = b^*(y \mid a)$.

So it can happen that $L$ does not $d$-separate $A$ and $Y$ in $G_{\overline{A}}$ in the causal graph $G$, while the statistical criteria holds. In that case we still have $b^*(y \mid a) = P(Y_a = y)$. for the effect of $A$ on $Y$. These examples involve perfect cancellations and are therefore not practically relevant. The statistical criteria for $b(y \mid a) = b^*(y \mid a)$ can be tested based on data, though.

# DEFINING NON-CONFOUNDING IN A CAUSAL GRAPH

**Graphical conditions for non-confounding in a causal graph.** Suppose that the graph is causal. If there is no back door path from $A$ to $Y$, then the effect of $A$ on $Y$ is NOT confounded.

If there is no back door path from $A$ to $Y$ controlling for $L$ (i.e. $Y$ is $d$-separated from $A$ by $L$ in $G_{\overline{A}}$), then the effect of $A$ on $Y$ within stata of $L$ is unconfounded and we call $L$ **sufficient set for adjustment**.

Thus if $L$ is a sufficient set for adjustment for the effect of $A$ on $Y$, then the $G$-computation formula $b^*(y \mid a)$ only adjusting for $L$ equals $P(Y_a = y)$. Thus in this case one can estimate the counterfactual distribution of $Y_a$

if one measures $L$ (the other potential confounders $U$ do not need to be measured).

Thus if one is able to provide a causal graph before planning a study to determine a (adjusted) causal effect of $A$ on $Y$, then one can use this causal graph to determine which variables need to be measured beyond $(A, Y)$.

# UNNECESSARY ADJUSTMENT

Consider a causal graph.

**Unnecessary adjustment and harmful adjustment:** One can have that the effect of $A$ on $Y$ is not confounded marginally (no back door path in $G_{\overline{A}}$), but that the effect of $A$ on $Y$, within strata $C$, is confounded.

See Figure 5.

LESSON: Adjustment for variables (such as C in Fig 5) that are not necessary to control may necessitate adjustment for even more variables, and there might not be anymore that would remove the bias (see Figure 6).

As a consequence the following can happen: the marginal $G$-computation formula might represent the causal effect of $A$ on $Y$ (i.e. $P(Y_a = y)$) while the adjusted $G$-computation formula does NOT represent an adjusted causal effect (i.e. $P(Y_a = y \mid C)$)!

If one has the causal graph available, then one can prevent this to happen, but otherwise this is an actual risk.

To give a concrete example: the data is $E, D, F$ and the true causal graph is Fig 6 which we do not know. Our goal is too estimate the marginal causal effect of $E$ on $D$. Suppose we worry about $F$ being a confounder and therefore we use the $G$-computation formula adjusting for $F$ (WRONG), while we could have used the marginal $G$-computation formula (CORRECT).

EXAMPLE 1 of adjustment induced bias: In studies of estrogen (E) and endometrical cancer (D), some researchers attempted to control for detection bias by stratifying on uterine bleeding (F), which could be caused by either estrogen or cancer, as in Figure 6. The association between estrogen and cancer withing levels of bleeding was drastically reduced by this stratification (likely due to bias produced by the adjustment).

EXAMPLE 2 (Healthy worker survivor effect): Unmeasured health conditions influence decision to leave work. Then leaving work is associated with mortality, even when it has no causal effect on mortality. Let the exposure (E) be job-assignment, which influences worker decisions to leave work (L). Fig 7 is the causal graph for this scenario.

The effect of E on D is marginally unconfounded but within strata of $L$ the effect of E on D is confounded.

# MINIMAL SUFFICIENT SET FOR ADJUSTMENT

A set $L$ is minimally sufficient for adjustment if $L$ is sufficient for adjustment, but no proper subset of $L$ is sufficient.
Fig 1: $\{A, C\}$ and $\{B, C\}$ are minimal sufficient.
Fig 5: $\{A, C\}$ and $\{B, C\}$ are sufficient, but not minimal sufficient.
To find a minimally sufficient set we may sequentially delete variables from a sufficient set until no more variables can be dropped without the new set failing the back door test (i.e. not being sufficient anymore).

Fact: $L$ can be sufficient while adding variables to $L$ can lead to an insufficient set.
Fig 5: $L = \{\}$ is sufficient, but $\{C\}$ is not sufficient.

Fact: There may exist several different minimal sufficient sets.
Fig 12: $\{A, B, C\}$ and $\{F\}$ are minimally sufficient sets of adjustment.

# IDENTIFIABILITY OF CAUSAL EFFECTS IN A CAUSAL GRAPH

Given a causal graph, suppose that one cannot find observed covariates $L$ so that $A$ and $Y$ are d-separated, given $L$. This does not imply that the causal effect of $A$ on $Y$ is not-identified, but there does not exist one standard formula such as the $G$-computation formula. The approach is the following. Let $(L, U)$ be a sufficient set for adjustment, i.e. $A \perp_d Y \mid (L, U)$, but the components $U$ will not be observable. Then we still have the $G$-computation formula (using Pearl's notation):

$$P(Y = y \mid \hat{a}) = \int P(Y = y \mid A = a, L = l, U = u) dF_{L,U}$$

The causal graph is a statistical graph and thus we have a special structure of the density of all nodes in the graph. Using the conditional independence assumptions of the statistical graph can sometimes be used to eliminate $U$ from the $G$-computation formula. This is a purely algebraic excercise. If one succeeds in doing this then one has proved that $P(Y = y \mid \hat{a})$ is still identifiable.

Pearl (1995) developes a "Calculus of Intervention" for causal graphs which can be helpful in carrying out this excercise.

**Theorem 3 (Pearl 1995)**

Rule 1 (insertion/deletion of observation):

$P(Y = y \mid \hat{a}, Z, W) = P(Y = y \mid \hat{a}, W)$ if $Y \perp_d Z \mid (A,$

Rule 2 (action/observation exchange):

$P(Y = y \mid \hat{a}, \hat{z}, W) = P(Y = y \mid \hat{a}, Z = z, W)$ if $Y \perp_d Z$

Rule 3 (insertion/deletion of actions):

$P(Y = y \mid \hat{a}, \hat{z}, W) = P(Y = y \mid \hat{a}, W)$ if $Y \perp_d Z \mid (A,$

where $Z(W)$ is the set of $Z$-nodes that are not ancestors of any $W$-node in $G_{\bar{X}}$.

With the help of this calculus one can prove the following theorem:

**Theorem.** (The front door criterion) Suppose a set of variables $Z$ satisfies the following conditions relative to an ordered pair of variables $(A, Y)$.: (i) $Z$ intercepts all directed paths from $A$ to $Y$, (ii) there is no back door path between $A$ and $Z$, and (iii) every back door path between $Z$ and $Y$ is blocked by $A$. Then the causal effect of $A$ on $Y$ is identifiable and given by:

$$P(Y = y \mid \hat{a}) = \sum_{a} P(Z = z \mid A = a) \sum_{a'} P(Y = y \mid A =$$

Consider Figure 3 of Pearl 1995.

# IDENTIFIABILITY OF CAUSAL EFFECTS IN A CAUSAL GRAPH

Given a causal graph, suppose that one cannot find observed covariates $L$ so that $A$ and $Y$ are d-separated, given $L$. This does not imply that the causal effect of $A$ on $Y$ is not-identified, but there does not exist one standard formula such as the $G$-computation formula. The approach is the following. Let $(L, U)$ be a sufficient set for adjustment, i.e. $A \perp_d Y \mid (L, U)$, but the components $U$ will not be observable. Then we still have the $G$-computation formula (using Pearl's notation):

$$P(Y = y \mid \hat{a}) = \int P(Y = y \mid A = a, L = l, U = u) dF_{L,U}$$

The causal graph is a statistical graph and thus we have a special structure of the density of all nodes in the graph. Using the conditional independence assumptions of the statistical graph can sometimes be used to eliminate $U$ from the $G$-computation formula. This is a purely algebraic excercise. If one succeeds in doing this then one has proved that $P(Y = y \mid \hat{a})$ is still identifiable.

Pearl (1995) developes a "Calculus of Intervention" for causal graphs which can be helpful in carrying out this excercise.

**Theorem 3 (Pearl 1995)**

Rule 1 (insertion/deletion of observation):

$$P(Y = y \mid \hat{a}, Z, W) = P(Y = y \mid \hat{a}, W)$$

if $Y \perp_d Z \mid (A, W)$ in $G_{\bar{A}}$.

Rule 2 (action/observation exchange):

$$P(Y = y \mid \hat{a}, \hat{z}, W) = P(Y = y \mid \hat{a}, Z = z, W)$$

if $Y \perp_d Z \mid (A, W)$ in $G_{\bar{A}\underline{Z}}$.

Rule 3 (insertion/deletion of actions):

$$P(Y = y \mid \hat{a}, \hat{z}, W) = P(Y = y \mid \hat{a}, W)$$

if $Y \perp_d Z \mid (A, W)$ in $G_{\bar{A}\overline{Z(W)}}$, where $Z(W)$ is the set of $Z$-nodes that are not ancestors of any $W$-node in $G_{\bar{A}}$.

With the help of this calculus one can prove the following theorem:

**Theorem.** (The front door criterion) Suppose a set of variables $Z$ satisfies the following conditions relative to an ordered pair of variables $(A, Y)$:

(i) $Z$ intercepts all directed paths from $A$ to $Y$, (ii) there is no back door path between $A$ and $Z$, and

(iii) every back door path between $Z$ and $Y$ is blocked by $A$. Then the causal effect $P(Y = y \mid \hat{a})$ of $A$ on $Y$ is identifiable and given by:

$$\sum_a P(Z = z \mid a) \sum_{a'} P(Y = y \mid a', z) P(A = a').$$

Consider Figure 3 of Pearl 1995.

**Example:** This graphical criterion permits identification of causal effects by measuring variables that are affected by treatment. Let $A$ be smoking, $Y$ lung cancer and $Z$ the amount of tar deposited in subject's lungs, $U$ are unmeasured confounders of the effect of smooking.

# Proof of front door criterion.

**Task 1**: $P(Z = z \mid \hat{x}) = P(Z = z \mid x)$ using rule 2.

**Task 2:** Compute $P(Y = y \mid \hat{z})$.

$$P(Y = y \mid \hat{z}) = \sum_x P(Y = y \mid X = x, \hat{z}) P(X = x \mid \hat{z}).$$

By rule 3: $P(X = x \mid \hat{z}) = P(X = x)$ (i.e. manipulating $Z$ has no effect on $X$ because $Z$ is a descendant of $X$ in $G$.) By rule 2:

$$P(Y = y \mid X = x, \hat{z}) = P(Y = y \mid X = x, Z = z)$$

if $Z \perp_d Y \mid X$ in $G_{\underline{Z}}$. Thus we conclude:

$$\begin{aligned} P(Y = y \mid \hat{z}) &= \sum_x P(Y = y \mid X = x, z) P(X = x) \\ &= E_X P(Y = y \mid X, Z = z). \end{aligned}$$

**Task 3**: Compute $P(Y = y \mid \hat{x})$. We have:

$$P(Y = y \mid \hat{x}) = \sum_z P(Y = y \mid Z = z, \hat{x}) P(Z = z \mid \hat{x})$$

$$= \sum_z P(Y = y \mid Z = z, \hat{x}) P(Z = z \mid X = x).$$

By rule 2

$$P(Y = y \mid Z = z, \hat{x}) = P(Y = y \mid \hat{z}, \hat{x})$$

since $Y \perp_d Z \mid X$ in $G_{\bar{X}\underline{Z}}$. By rule 3 we have:

$$P(Y = y \mid \hat{z}, \hat{x}) = P(Y = y \mid \hat{z})$$

since $Y \perp_d X \mid Z$ in $G_{\overline{XZ}}$. Thus we have:

$$P(Y = y \mid Z = z, \hat{x}) = P(Y = y \mid \hat{z}).$$

In task 2 we already calculated $P(Y = y \mid \hat{z})$. Thus we have shown $P(Y = y \mid \hat{x})$ equals

$$\sum_{z} P(Z = z \mid x) \sum_{x'} P(Y = y \mid x', z) P(X = x').$$

# CAUSAL INFERENCE BY SURROGATE EXPERIMENTS

Suppose we wish to learn the causal effect of $A$ on $Y$ when $P(y \mid \hat{a})$ is not identifiable (due to unmeasured confounders) and for practical (ethical) reasons we cannot randomize $A$. Can we identify $P(y \mid \hat{a})$ by randomizing a surrogate variable $Z$ which is easier to control than $A$. For example, $A$ is cholesterol level, $Y$ is heart disease and $Z$ is diet.

**Theorem:** If (i) $A$ intercepts all directed paths form $Z$ to $Y$ and (ii) $P(Y \mid \hat{a})$ is identifiable in $G_{\bar{Z}}$ (the causal graph in which all incoming arrows in $Z$ are deleted).

**Proof.** If (i) holds, we have $P(y \mid \hat{a}) = P(y \mid \hat{a}, \hat{z})$ since $Y \perp_d Z \mid A$ in $G_{\bar{A}\bar{Z}}$. $P(y \mid \hat{a}, \hat{z})$ is the causal effect of $A$ on $Y$ in the causal graph $G_{\bar{Z}}$ which is identifiable by (ii). $\square$

Translated to our cholesterol example, there should be no direct effect of diet on heart disease and no confouding effect between cholesterol and heart disease, unless we can measure an intermediate variable between the two.

See figures 7e and 7h????

# PART III: G-COMPUTATION FORMULA

# G-COMPUTATION IN LONGITUDINAL STU

Let $A(j)$ be treatment assigned at time $j$, $L(j)$ covariate values measured after $A(j-1)$ and before $A(j)$, $j=0,\ldots,K$. Let $Y=L_{K+1}$ be the outcome of interest. Then the temporal ordering of all measured variables is given by:

$$L(0),A(0),L(1),\ldots,L(K),A(K),Y=L(K{+}1).$$

**Meaning of temporal ordering:** The future variables cannot affect the past variables: e.g. the counterfactual $L(0)_{A(0)=a(0)}$ is not affected by $a(0)$.

The corresponding density representation is given by:

$$f(v)=f(l_0)f(a_0\mid l_0)f(l_1\mid a_0,l_0)\ldots f(l_{K+1}\mid \bar{l}_K,\bar{a}_K).$$

Given a treatment vector $\overline{a}^*$, the density $f_{\overline{a}^*}(v)=f_{\overline{a}^*}(y,\overline{l}_K)$ is defined by the density $f(v)$ except that $f(a_j \mid \bar{a}_{j-1}, \bar{l}_j)$ is replaced by a degenerate distribution at $a_j^*$.

By integrating out $\bar{l}_K$ in this joint density $f_{\bar{a}^*}(v)$ we can obtain the marginal density $f_{\bar{a}^*}(y)$:

$$\int \ldots \int f(y \mid \bar{l}_K, \bar{a}_K^*) \prod_{j=1}^{K} f(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}^*) d\mu(l_j).$$

Thus the marginal distribution $F_{\bar{a}^*}$ is given by:

$$\int \ldots \int P(Y < y \mid \bar{l}_K, \bar{a}_K^*) \prod_{j=1}^{K} f(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}^*) d\mu(l_j).$$

Robins refers to this as the $G$-computation algorithm formula or functional for the effect of treatment action $\bar{A} = \bar{a}^*$ on the outcome $Y$. If the statistical graph is causal or if treatment assignment of $A(j)$ is sequentially randomized then

$$F_{\bar{a}^*}(y) = P(Y_{\bar{a}^*} \leq y).$$

Let's state this as a theorem.

**Theorem.** Suppose the ordering

$L(0), A(0), L(1), \ldots, L(K), A(K), Y = L(K{+}1)$

is temporal in the sense that $L(j)$ is only affected by $\bar{A}(j-1)$ for $j = 1, \ldots, K+1$. Consider the $G$-computations formulas $f_{\bar{a}^*}(y, \bar{l}_K)$ and $f_{\bar{a}^*}(y)$ corresponding with this ordering.

If

$$A(j) \perp (Y_{\bar{a}}, L_{\bar{a}} : \bar{a} \in \mathcal{A}) \mid \bar{L}(j), \bar{A}(j-1),$$

then the $G$-computation formula $f_{\bar{a}^*}(y, \bar{l}_K)$ equals

$$P(Y_{\bar{a}^*} = y, \bar{L}_{K,\bar{a}^*} = \bar{l}_K).$$

If

$$A(j) \perp (Y_{\bar{a}} : \bar{a} \in \mathcal{A}) \mid \bar{L}(j), \bar{A}(j-1),$$

then the $G$-computation formula $f_{\bar{a}^*}(y$ equals

$$P(Y_{\bar{a}^*} = y).$$

**Proof.** Give the general proof, see handout (Maja).

# JAMIE'S HYPOTHETICAL EXAMPLE

Let $A_0$ be a randomly assigned treatment (drugs, yes or no) assigned at $t_0$, $L$ is indicator of having developed a risk factor such as Pneumonia at time $t_1$, $A_1$ is treatment (AZT) indicator at time $t_1$ (which can be based on values of $A_0, L$) and $Y$ is an outcome at $t_2$ such as the indicator of being alive at $t_2$. In this example, we can think of $A_0 = 1$ as a drug which prevents the development of Pneumonia ($L = 1$).

**Question 1:** Estimate causal effect of $A_0$. In other words, estimate $P(Y_{A_0=1} = 1) - P(Y_{A_0=0} = 1)$, where $Y_{A_0=0}$ ($Y_{A_0=0} = 1$) is the counterfactual outcome we would have observed on everybody if everybody gets assigned $A_0 = 0$.

**Answer:** $P(Y = 1 \mid A_0 = 1) - P(Y = 1 \mid A_0 = 0) = 8/16 - 10/16 = -1/8$. So marginally treating hurts.

**Question 2:** Would it have been wrong to adjust for $L$ in Question 1? In other words, would

$$P(Y = 1 \mid A_0 = 1, L = 1) - P(Y = 1 \mid A_0 = 0, L = 1)$$

have a causal interpretation.
**Answer:** If a subject developes Pneumonia $(L = 1)$ *in spite of* treatment $A_0 = 1$, then that says something extra about the subject relative to a subject who developed Pneumonia $(L = 1)$ in the control treatment arm $A_0 = 0$. Formally, since $L = L_{A_0}$ the conditioning event $A_0 = 1, L = 1$ equals $A_0 = 1, L_1 = 1$ while the conditioning event $A_0 = 0, L = 1$ equals $A_0 = 0, L_0 = 1$

**Question 3:** Suppose that we would like to know which of the two treatment regimes $A_0 =$

$0, A_1 = 1$ and $A_0 = 1, A_1 = 1$ are best. Then we want to estimate $P(Y_{11} = 1) - P(Y_{01} = 1)$. How?

NAIVE I: $P(Y = 1 \mid A_0 = 0, A_1 = 1) - P(Y = 1 \mid A_0 = 1, A_1 = 1)$. Wrong since $L$ is a confounder of $A_1$ and $A_0$ affects $L$.
NAIVE II: Adjust for $L$: $P(Y = 1 \mid A_0 = 0, A_1 = 1, L = 1) - P(Y = 1 \mid A_0 = 1, A_1 = 1, L = 1)$. In the example, this difference equals 1/8 indicating treating at $t_0$ hurts in the $L = 1$ strata.
Wrong, cannot adjust for covariate affected by treatment. As above, having $L = 1$ in $A_0 = 1$ group is a very different statement from having $L = 1$ in $A_0 = 0$ group.

G-COMPUTATION FORMULA:

$$P_{a_0 a_1}(Y = 1, L = l) = P(L = l)P(Y = 1 \mid a_0, a_1, l).$$

Thus $P(Y_{a_0 a_1} = 1)$ is given by:

$$P_{a_0 a_1}(Y = 1, L = 1) + P_{a_0 a_1}(Y = 1, L = 0).$$

This formula gives:

$$P(Y_{11} = 1) = 1/2 * 1/2 + 1/2 * 3/4 = 5/8.$$

And

$$P(Y_{01} = 1) = 1 * 10/16 + 0 = 5/8.$$

Note that $P(Y_{00} = 1)$ and $P(Y_{10} = 1)$ are not identified from data example.

# ALTERNATIVE REPRESENTATION OF G-COMP FORMULA

Recall the $G$-comp formula:

$$f_{\bar{a}^*}(y) = \int \dots \int f(y \mid \bar{l}_K, \bar{a}^*_K) \prod_{j=1}^{K} f(l_j \mid \bar{l}_{j-1}, \bar{a}^*_{j-1}) d\mu(l_j$$

This can be rewritten as:

$$E\left(\frac{I(Y = y, \bar{A} = \bar{a}^*)}{\prod_{j=0}^{K} P(A(j) = a^*(j) \mid \bar{A}_{j-1} = \bar{a}_{j-1}, \bar{L}_{j-1})}\right).$$

# G-COMPUTATION FORMULA

Consider a statistical graph for a set of nodes $X_1, \ldots, X_m$, where we will assume that these are ordered temporarily. Then the corresponding representation of the density of $X_1, \ldots, X_m$ is given by:

$$p(x_1, \ldots, x_m) = \prod_{i=1}^{m} P(X_i = x_i \mid (PA)_i = (pa)_i),$$

where $P(X_i = x_i \mid (PA)_i = (pa)_i)$ is the conditional density of $X_i$, given its parents $(PA)_i$ in the statistical graph.

Let $A$ be a subset of the nodes $(X_1, \ldots, X_m)$. Let's denote the remainder of the nodes with $(Y, W)$, where $Y$ is an outcome variable of interest. The $G$-computational formula for the effect of $A = a$ on $Y$ is a functional of this joint density representation: so it depends on the ordering of the variables as well.

**How would you obtain from this joint density the density of $(Y, W)$ in the hypothetical world where we set $A = a$:** Suppose that the statistical graph is even causal. Then we can represent the world of $(X_1, \ldots, X_m)$ by a system of $m$ equations $X_i = \phi_i((PA)_i, \epsilon_i)$, $i = 1, \ldots, m$. What is the distribution of the variables $(Y, W)$ if we set $A = a$ in this system? Setting $A = a$ just reduces the number of equations since all equations corresponding with $X_i \in A$ are deleted and we set $A = a$ in all other equations. This is just a new causal graph and thus we can write down its corresponding density.

If we **set** $A = a$ (i.e. we intervene by setting $A = a$, but otherwise remain things as they are), then a new density $p(Y = y, W = w \mid \hat{a})$ of the graph is obtained by setting the conditional densities of nodes in $A$ equal to a degenerate density at $A = a$:

$$\frac{\prod_{i=1}^{m} P(X_i = x_i \mid (PA)_i = (pa)_i)}{\prod_{X_i \in A} P(X_i = x_i \mid (PA)_i = (pa)_i)}$$

and this object is evaluated at $(x_1, \ldots, x_m)$ corresponding with $(y, w, a)$. This density represents the density of $(Y, W)$ in the hypothetical world where we set $A = a$.

Suppose we want to obtain a formula for the causal effect of setting $A = a$ on an outcome variable $Y$, where $Y$ is one of the nodes. Then we find this by integrating out all other variables in $p(Y = y, W = w \mid \hat{a})$:

$$b(y \mid a) = \int_w P(Y = y, dw \mid \hat{a}).$$

This is the $G$-computation formula of Robins. If the graph is causal, then this equals $P(Y_a = y)$. More general, if the necessary (sequential) randomization assumption holds for the data $(A, Y, W)$ (i.e. $(X_1, \ldots, X_n)$), then this equals $P(Y_a = y)$.

## G-COMPUTATION IN LONGITUDINAL STU

Let $A(j)$ be treatment assigned at time $j$, $L(j)$ covariate values measured after $A(j-1)$ and before $A(j)$, $j = 0, \ldots, K$. Let $Y = L_{K+1}$ be the outcome of interest. Then the temporal ordering of all measured variables is given by:

$$L(0), A(0), L(1), \ldots, L(K), A(K), Y = L(K+1).$$

**Meaning of temporal ordering:** The future variables cannot affect the past variables: e.g. the counterfactual $L(0)_{A(0)=a(0)}$ is not affected by $a(0)$.

The corresponding density representation is given by:

$$f(v) = f(l_0)f(a_0 \mid l_0)f(l_1 \mid a_0, l_0)\ldots f(l_{K+1} \mid \bar{l}_K, \bar{a}_K).$$

Given a treatment vector $\bar{a}^*$, the density $f_{\bar{a}^*}(v) = f_{\bar{a}^*}(y, \bar{l}_K)$ is defined by the density $f(v)$ except that $f(a_j \mid \bar{a}_{j-1}, \bar{l}_j)$ is replaced by a degenerate distribution at $a_j^*$.

By integrating out $\bar{l}_K$ in this joint density $f_{\bar{a}^*}(v)$ we can obtain the marginal density $f_{\bar{a}^*}(y)$:

$$f_{\bar{a}^*}(y) = \int \dots \int f(y \mid \bar{l}_K, \bar{a}_K^*) \prod_{j=1}^{K} f(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}^*) d\mu(l_j$$

Thus the marginal distribution $F_{\bar{a}^*}$ is given by:

$$F_{\bar{a}^*}(y) = \int \dots \int P(Y < y \mid \bar{l}_K, \bar{a}_K^*) \prod_{j=1}^{K} f(l_j \mid \bar{l}_{j-1}, \bar{a}_{j-1}^*)$$

Robins refers to this as the $G$-computation algorithm formula or functional for the effect of treatment action $\bar{A} = \bar{a}^*$ on the outcome $Y$. If the statistical graph is causal or if treatment assignment of $A(j)$ is sequentially randomized then

$$F_{\bar{a}^*}(y) = P(Y_{\bar{a}^*} \leq y).$$

**Theorem.** Suppose the ordering

$L(0), A(0), L(1), \ldots, L(K), A(K), Y = L(K{+}1)$

is temporal in the sense that $L(j)$ is only affected by $\bar{A}(j-1)$ for $j = 1, \ldots, K+1$. Consider the $G$-computations formulas $f_{\bar{a}^*}(y, \bar{l}_K)$ and $f_{\bar{a}^*}(y)$ corresponding with this ordering.

If

$$A(j) \perp (Y_{\bar{a}}, L_{\bar{a}} : \bar{a} \in \mathcal{A}) \mid \bar{L}(j), \bar{A}(j-1),$$

then the $G$-computation formula $f_{\bar{a}^*}(y, \bar{l}_K)$ equals

$$P(Y_{\bar{a}^*} = y, \bar{L}_{K,\bar{a}^*} = \bar{l}_K).$$

If

$$A(j) \perp (Y_{\bar{a}} : \bar{a} \in \mathcal{A}) \mid \bar{L}(j), \bar{A}(j-1),$$

then the $G$-computation formula $f_{\bar{a}^*}(y$ equals

$$P(Y_{\bar{a}^*} = y).$$

**Proof.** For simplicity: give the proof for $L_0, A_0, L_1, A_1, Y$.

## G-COMPUTATION IN SIMPLE EXAMPLE.

Suppose that the data on a subject is $(A, Y, W_1, W_2)$, where $A$ is treatment, $Y$ is outcome, $W_1, W_2$ are covariates. Assume the following temporal ordering at which the variables are generated:

$$W = (W_1, W_2), A, Y.$$

In other words, one first generates covariates, then the treatment is drawn possibly based on $W$ and subsequently one measures the outcome $Y$.

Determine the $G$-computation formula for: $P(Y_a \leq y)$ and $P(Y_a \leq y \mid W_1)$.

# JAMIE'S HYPOTHETICAL EXAMPLE

Let $A_0$ be a randomly assigned treatment (yes or no) assigned at $t_0$, $L_1$ is indicator of having developed a risk factor such as Anemia at time $t_1$, $A_1$ is treatment (AZT) indicator at time $t_1$ (which can be based on values of $A_0, L_1$) and $Y$ is an outcome at $t_2$ such as the indicator of being alive at $t_2$. In this example, we can think of $A_0 = 1$ as a treatment which prevents the development of Anemia ($L_1 = 1$).

**Question 1:** Estimate causal effect of $A_0$. In other words, estimate $P(Y_{A_0=1} = 1) - P(Y_{A_0=0} = 1)$, where $Y_{A_0=0}$ $(Y_{A_0=0} = 1)$ is the counterfactual outcome we would have observed on everybody if everybody gets assigned $A_0 = 0$.

**Answer:** $P(Y = 1 \mid A_0 = 1) - P(Y = 1 \mid A_0 = 0) = 8/16 - 10/16 = -1/8$. So marginally treating hurts.

**Question 2:** Would it have been wrong to adjust for $L_1$ in Question 1? In other words, would

$$P(Y = 1 \mid A_0 = 1, L_1 = 1) - P(Y = 1 \mid A_0 = 0, L_1 = 1$$

have a causal interpretation.
**Answer:** If a subject developes Anemia ($L_1 = 1$) *in spite of* treatment $A_0 = 1$, then that says something extra about the subject relative to a subject who developed Anemia ($L_1 = 1$) in the control treatment arm $A_0 = 0$. So an association between $A_0$ and $Y$ in the group $L_1 = 1$ can be solely due to the fact that $A_0 = 1$ prevents $L_1 = 1$.

**Question 3:** Suppose that we would like to know which of the two treatment regimes $A_0 = 0, A_1 = 1$ and $A_0 = 1, A_1 = 1$ are best. Then we want to estimate $P(Y_{11} = 1) - P(Y_{01} = 1)$. How?

NAIVE I: $P(Y = 1 \mid A_0 = 0, A_1 = 1) - P(Y = 1 \mid A_0 = 1, A_1 = 1)$. Wrong since $L_1$ is a confounder of $A_1$ and $A_0$ affects $L_1$.
NAIVE II: Adjust for $L_1$: $P(Y = 1 \mid A_0 = 0, A_1 = 1, L_1 = 1) - P(Y = 1 \mid A_0 = 1, A_1 = 1, L_1 = 1)$. In the example, this difference equals 1/8 indicating treating at $t_0$ hurts in the $L_1 = 1$ strata.
Wrong, cannot adjust for covariate affected by treatment. Having $L_1 = 1$ in $A_0 = 1$ group is a very different statement from having $L_1 = 1$ in $A_0 = 0$ group.

G-COMPUTATION FORMULA:

$P_{a_0 a_1}(Y = 1, L_1 = l_1) = P(L_1 = l_1) P(Y = 1 \mid A_0 = a$

Thus $P(Y_{a_0a_1} = 1)$ is given by:

$P_{a_0a_1}(Y = 1, L_1 = 1) + P_{a_0a_1}(Y = 1, L_1 = 0)$.

This formula gives:

$$P(Y_{11} = 1) = 1/2 * 1/2 + 1/2 * 3/4 = 5/8.$$

And

$$P(Y_{01} = 1) = 1 * 10/16 + 0 = 5/8.$$

Note that $P(Y_{00} = 1)$ and $P(Y_{10} = 1)$ are not identified from data example.

# ALTERNATIVE REPRESENTATION OF G-COMP FORMULA

Recall the $G$-comp formula:

$$f_{\bar{a}^*}(y) = \int \dots \int f(y \mid \bar{l}_K, \bar{a}^*_K) \prod_{j=1}^{K} f(l_j \mid \bar{l}_{j-1}, \bar{a}^*_{j-1}) d\mu(l_j$$

This can be rewritten as:

$$f_{\bar{a}^*}(y) = E \frac{I(Y = y, \bar{A} = \bar{a}^*)}{\prod_{j=0}^{K} P(A(j) = a^*(j) \mid \bar{A}_{j-1} = \bar{a}_{j-1}, \bar{L}_{j-1})}.$$

yes

# PART III: MARGINAL STRUCTURAL MODELS. IN POINT TREATMENT STUDIES

# REGRESSION MODELS.

Consider the regression model:

$$Y = m_\alpha(A, V) + \epsilon, \;\; E(\epsilon \mid A, V) = 0,$$

where $m_\alpha(A, V) = E(Y \mid A, V)$ is a given parametrization of the regression surface. The observed data is $n$ observations on $(Y, A, W)$, where $V$ is a subset of the observed covariates $W$, and the goal is to estimate $\alpha \in \mathbf{R}^k$. Here $A$ is a treatment variable, $W$ are covariates and $Y$ is an outcome variable of interest. Let $\epsilon(\alpha) \equiv Y - m_\alpha(A, V)$.

**EXAMPLE:** If $Y$ is Bernoulli one could assume:

$$\begin{aligned} P(Y = 1 \mid A, V) &= \alpha_0 + \alpha_1 A + \alpha_2 V \\ P(Y = 1 \mid A, V) &= \exp(\alpha_0 + \alpha_1 A + \alpha_2 V) \\ P(Y = 1 \mid A, V) &= \frac{1}{1 + \exp(\alpha_0 + \alpha_1 A + \alpha_2 V)} \end{aligned}$$

In these three models the parameter $\alpha_1$ represents the (adjusted) *Risk Difference*, *Relative Risk* and the *Odds Ratio*, respectively.

Each vector function $(A, V) \to h(A, V) \in \mathbf{R}^k$ implies an unbiased estimating equation for $\alpha$ given by:

$$0 = \sum_{i=1}^{n} h(A_i, V_i)\epsilon_i(\alpha).$$

(Note that the least squares estimator would correspond with $h(A, V) = d/d\alpha m_\alpha(A, V)$.)

Under weak regularity conditions we have that the solution $\alpha_n$ is root-$n$ consistent and asymptotically linear:

$$\sqrt{n}(\alpha_n - \alpha) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} C^{-1} h(A_i, V_i)\epsilon_i(\alpha) + o_P(1),$$

where the $k \times k$-matrix $C$ is given by:

$$C = E\left\{h(A,V)\frac{d}{d\alpha}m_\alpha(A,V)\right\}.$$

In other words, with $X = (A, W, Y)$ we have

$$\sqrt{n}(\alpha_n - \alpha) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC(X_i \mid \alpha)$$

where $IC(X \mid \alpha)$ is the so called influence curve given by:

$$IC(X \mid \alpha) \equiv C^{-1}h(A,V)\epsilon(\alpha).$$

**Global summary of Proof:** Let $X = (Y, A, W)$. Define $S_\alpha(X) = h(A,V)\epsilon(\alpha)$, let $\alpha_0$ be the true regression parameter, $P_0$ be the true data generating distribution and let $P_n$ be the empirical distribution of the data. Assume that we have shown consistency of $\alpha_n$ by other means. We have:

$$E_{P_0}\{S_{\alpha_n} - S_{\alpha_0}\} = -E_{P_n-P_0}S_{\alpha_n}(X).$$

Empirical process theory shows that:

$$E_{P_n - P_0} S_{\alpha_n}(X) = E_{P_n - P_0} S_{\alpha_0}(X) = o_P(1/\sqrt{n}).$$

The latter can be written as:

$$\frac{1}{n} \sum_{i=1}^{n} S_{\alpha_0}(X_i).$$

If we have differentiability, then

$$\begin{aligned} E_{P_0}\{S_{\alpha_n} - S_{\alpha_0}\} &= \frac{d}{d\alpha} E_{P_0} S_\alpha(X) \Big|_{\alpha=\alpha_0} (\alpha_n - \alpha_0) \\ &\quad + o(\mid \alpha_n - \alpha_0 \mid). \end{aligned}$$

Let

$$C \equiv \frac{d}{d\alpha} E_{P_0} S_\alpha(X) \Big|_{\alpha=\alpha_0}.$$

Then we have:

$$C(\alpha_n - \alpha_0) = \frac{1}{n} \sum_{i=1}^{n} S_{\alpha_0}(X_i) + oP(1/\sqrt{n}).$$

Applying $C^{-1}$ to both sides gives the wished results. □

Thus (by central limit theorem) $\sqrt{n}(\alpha_n - \alpha)$

is asymptotically normally distrubuted. The normal limit distribution has expectation zero and covariance matrix given by:

$$\begin{aligned}\Sigma &= E(IC(X \mid \alpha)IC(X \mid \alpha)^\top) \\ &= C^{-1}E\{h(A,V)h(A,V)^\top \epsilon(\alpha)^2\}C^{-1}.\end{aligned}$$

Given an estimator $\alpha_n$ of $\alpha$ we can estimate $\Sigma$ with the empirical covariance matrix of $IC(X_i \mid \alpha_n)$, $i = 1, \ldots, n$. This can be used to construct an asymptotic 0.95 confidence interval for each component $\alpha_j$ of $\alpha$.

The optimal covariance matrix (smallest variance on the diagonal) is obtained by setting

$$h = h_{opt}(A,V) = \frac{\frac{d}{d\alpha}m_\alpha(A,V)}{E(\epsilon^2(\alpha) \mid A,V)}.$$

The solution of $0 = \sum_{i=1}^n h_{opt}(A_i,V_i)\epsilon_i(\alpha)$ equals the following weighted least squares estimator:

$$\alpha_{n,opt} = \mathsf{min}^{-1} \sum_{i=1}^{n} w_i \{Y_i - m_\alpha(A_i, V_i)\}^2 \,,$$

where

$$w_i = \frac{1}{\mathsf{E}(\epsilon^2(\alpha) \mid A_i, V_i)}.$$

This estimator is not available in practice since the weights are unknown. However, it immediately suggests an iterative weighted least squares estimator: HOW, Describe it in detail.

This iterative weighted least squares estimator (IWLSE) requires guessing a model for the regression $E(\epsilon^2(\alpha) \mid A, V)$. If this guessed model is correct, then the resulting IWLSE is asymptotically efficient. If the guessed model is wrong, then the resulting IWLSE is still consistent and asymptotically normal. Therefore we call this IWLSE estimator a locally efficient estimator of $\alpha$ at the guessed model.

# A CAUSAL REGRESSION MODEL FOR POINT TREATMENT

Let $A$ be a treatment variable with outcome space $\mathcal{A}$, $W$ be a vector of baseline covariates not affected by $A$ and $Y$ is an outcome variable. Define the vector of treatment specific counterfactuals $(Y_a : a \in \mathcal{A})$. Assume that $A$ is randomized w.r.t. $W$:

$$P(A = a \mid (Y_a : a \in \mathcal{A}), W) = P(A = a \mid W).$$

We will denote the latter propensity score with $g(a \mid W)$.

We assume the following causal regression model: for each $a \in \mathcal{A}$

$$E(Y_a \mid V) = m_\beta(a, V) + \epsilon_a,$$

where $E(\epsilon_a \mid V) = 0$. Such a model is called a *Marginal Structural Model*. Note that $\beta$ is a causally interpretable parameter.

**EXAMPLE:** If $Y$ is Bernoulli one could assume:

$$\begin{aligned} P(Y_a = 1 \mid V) &= \beta_0 + \beta_1 a + \beta_2 V \\ P(Y_a = 1 \mid V) &= \exp(\beta_0 + \beta_1 a + \beta_2 V) \\ P(Y_a = 1 \mid V) &= \frac{1}{1 + \exp(\beta_0 + \beta_1 a + \beta_2 V)} \end{aligned}$$

In these three models the parameter $\beta_1$ represents the (adjusted) *Causal Risk Difference*, *Causal Relative Risk* and the *Causal Odds Ratio*, respectively.

**When does $\alpha$ equal $\beta$.** In other words, when do we have $E(Y \mid A = a, V) = E(Y_a \mid V)$? Answer: if $A$ is randomized w.r.t. $V$ (i.e. $A$ is completely selected at random within stata of $V$). Formally,

$$g(a \mid (Y_a : a \in \mathcal{A}), W) == g(a \mid V).$$

In that case, we have that $\alpha_1$ represents the causal effect of $A$ on $Y$ within strata of $V$. This requires adjusting for all potential confounders in the regression model. In that case we have a locally efficient estimator for $\alpha$, as given above, and thus of $\beta$ (since $\alpha = \beta$).

## ESTIMATING EQUATIONS FOR $\beta$.

Each vector function $(A, V) \rightarrow h(A, V) \in \mathbf{R}^k$ implies an unbiased estimating equation for $\beta$ given by:

$$0 = \sum_{i=1}^{n} \frac{h(A_i, V_i)}{g(A_i \mid W_i)} \epsilon_{A_i}(\beta).$$

If

$(Condition)$ for almost every $W$ and each $a$ $h(a, V)/$

(1)

then one can indeed show

$$E\left\{\frac{h(A, V)}{g(A \mid W)} \epsilon_A(\beta)\right\} = 0.$$

Give the proof.

Discuss this identifiability condition. Firstly, we note that this condition is needed to make the causal parameter identifiable from the data. Nonparametric estimation of the G-computation formula $E(Y_a \mid V) = EE(Y \mid A = a, W) \mid V)$ would require that the conditioning event $(A = a, W)$ always has positive probability. Therefore this condition should not come as a surprise. Before doing an analysis it is advisable to plot empirically $(A_i, W_i)$, $i = 1, \ldots, n$, in order to detect subpopulations $W = w$ for which $g(a \mid w) = 0$ for some $a$.

The fact that this condition depends on $h$ and thus on the choice of the estimating equation is helpful. For example, it might be possible to set $h(A, V) = h_1(A, V) I(A \in \mathcal{A}_1, V \in \mathcal{V}_1)$ for some subset $\mathcal{A}_1$ of all treatment outcomes and some subset $\mathcal{V}_1$ of covariate values for which $g(a \mid W) > 0$ for all $a \in \mathcal{A}_1$, $V \in \mathcal{V}_1$.

Consider now the scenario in which subjects with a certain covariate value $W = w$ always receive treatment 1. Then it would make most sense to delete these subjects from the sample. One will now do causal inference for the population of subjects with $W \neq w$, which supposedly is the population of interest since doctors already knew the best treatment for subjects with $W = w$. However, in case one is truly interested in doing causal inference for the total population one could model and estimate $E(Y \mid A, W)$ (which thus involves extrapolating this surface to the region of $A, W$'s for which no data is available) and use the $G$-computation formula $E(Y_a \mid V) = EE(Y \mid A = a, W) \mid V)$. However, keep in mind that the consistency of the estimate relies on having guessed what the effect of the other treatments would have been for subjects with $W = w$.

**Back to the estimating equation:** Since $g(a \mid W)$ is an unknown nuisance parameter in this estimating equation this insights results in the following proposed estimators: for each $h(A, V)$ and an estimator $g_n(\cdot \mid W)$ of $g(\cdot \mid W)$ we have the following estimating equation:

$$0 = \sum_{i=1}^{n} \frac{h(A_i, V_i)}{g_n(A_i \mid W_i)} \epsilon_{A_i}(\beta).$$

We refer to these type of estimators of $\beta$ as the Inverse of Probability of Treatment Weighted (IPTW) estimator. We propose (Robins) to choose

$$h(A, V) = h^*(A, V) \equiv \frac{g(A \mid V) \frac{d}{d\beta} m_\beta(A, V)}{E(\epsilon_A^2(\beta) \mid A, V)}.$$

The advantages of this choice of estimating equation is:

1) If $A$ is randomized w.r.t. $V$, then this estimating equation corresponds with the estimating equation $0 = \sum_{i=1}^{n} h_{opt}(A_i, V_i)\epsilon_i(\beta)$ which is in this situtation the optimal estimating equation.

2) In general, $g(A \mid V)/g(A \mid W)$ is much more

stable than $1/g(A \mid W)$.
To summarize: multiplying with $g(A \mid V)$ stabilizes the estimating equation in general and it makes the estimating equation even optimal when all confounders are contained in $V$.

The solution of $0 = \sum_{i=1}^{n} \frac{h^*(A_i,V_i)}{g(A_i|W_i)} \epsilon_{A_i}(\beta)$ equals the following weighted least squares estimator:

$$\beta_n = \min^{-1} \sum_{i=1}^{n} w_i \left\{ Y_i - m_\beta(A_i, V_i) \right\}^2 ,$$

where

$$w_i = \frac{g(A_i \mid V_i)}{g(A_i \mid W_i) \mathrm{E}(\epsilon^2(\alpha) \mid A_i, V_i)}.$$

This estimator is not available in practice since $g(A \mid V), g(A \mid W)$ and $E(\epsilon^2(\beta) \mid A_i, V_i)$ are unknown. However, it immediately suggests an iterative weighted least squares estimator: HOW, Describe it in detail?

This iterative weighted least squares estimator of $\beta$ requires a choice of model for $g(A \mid W)$, $g(A \mid V)$ and for the regression of $\epsilon^2(\beta)$ on $A, V$. The model for $g(A \mid W)$ implies a model for $g(A \mid V)$: just assume that the regression parameters in front of the covariates beyond

$V$ are equal to zero. The consistency of the estimator $\beta_n$ only relies on consistent estimation of (i.e. the correct model for) $g(A \mid W)$ and on the correctness of the marginal structural model $E(Y_a \mid V) = m_\beta(a, V)$.

**Choices of models for the propensity score:**
**Bernoulli:** If $A$ is a bernoulli random variable, one can select a logistic regression model for $g(A \mid W)$.
**Discrete:** If $A$ is discrete, then one can use a multinomial regression:

$$\begin{aligned} P(A_0 = a_0 \mid W) &= \frac{\exp(\gamma_{a_0} + \gamma_1 W)}{1 + \sum_{a_0 \neq 0} \exp(\gamma_{a_0} + \gamma_1 W)} \\ P(A_0 = 0 \mid W) &= \frac{1}{1 + \sum_{a_0 \neq 0} \exp(\gamma_{a_0} + \gamma_1 W)}. \end{aligned}$$

Or Poisson regression:

$$P(A = a \mid W) = \frac{\lambda(W)^a}{a!} \exp(-\lambda(W)),$$

where we assume some regression model for

$\lambda(W) = E(A \mid W)$.

**Continuous:** If $A$ is a continuous variable, then

1) assume that $E(A \mid W) = m_\gamma(A, W)$ for some regression model $m_\gamma$ and that the error distribution follows a known family (e.g. normal error disribution) with possibly a few unknown parameters. The regression estimation is then standard and the residuals can then be used to fit the parametric error distribution.

2) One could also use a semiparametric model such as the Cox-proportional hazards model:

$$\lambda(a \mid W) = \lambda_0(a) \exp(\gamma W).$$

Or any other semiparametric model such as the accelerated failure time model, the very flexible HAAR hazard models of Stone and Kooperberg among many others.

**Important fact:** If one estimates $g(A \mid W)$ more nonparametrically, then the asymptotic efficiency of the estimator $\beta_n$ increases. Therefore one should choose the dimension of the model for $g(a \mid W)$ as large as sample size allows.

**Compare this IPTW-estimator $\beta_n$ with an estimator based on the $G$-computation formula.**

+ +

# PART IV:

# MARGINAL STRUCTURAL MODELSnl FOR TIME-DEPENDENT TREATMENT

Consider a longitudinal study with data collected in the following temporal ordering:

$$L_0, A_0, L_1, A_1, \ldots, L_K, A_K, Y.$$

Let $V$ be a subset of the baseline covariates $L_0$. Let $\bar{A}_k = (A_0, \ldots, A_k)$ be the treatment or exposure history up till time $k$ and $\bar{A} = (A_0, \ldots, A_k)$ is the treatment history up till end of follow up. Similarly, we define $\bar{L}_k$ and $\bar{L}$. For convenience, we will now and then use the notation $L_{K+1} = Y$.

Let $Y_{\bar{a}}$ be the counterfactual value of $Y$ that would have been observed had the subject received treatment history $\bar{a} = (a_0, \ldots, a_K)$. We can also define counterfactuals $L_{\bar{a}}$ which denotes the process $L$ that would have been observed if the subject had received treatment $\bar{a}$. The $Y_{\bar{a}}$, $\bar{a} \in \mathcal{A}$, are the counterfactuals of interest.

We will assume that treatment is sequentially randomized: for each possible treatment regime $\bar{a}$ (consistent with the observed history)

$$A(k) \perp Y_{\bar{a}} \mid \bar{A}(k-1), \bar{L}(k).$$

In other words, for each $k$

$$g(a(k) \mid (Y_{\bar{a}} : \bar{a} \in \mathcal{A}), \bar{A}(k-1), \bar{L}(k))$$

$$= g(A(k) \mid \bar{A}(k-1), \bar{L}(k)).$$

We define:

$$g(\bar{a} \mid X) = \prod_{k=0}^{K} g(a(k) \mid \bar{A}(k-1), \bar{L}(k)),$$

which one can think of as the conditional probability on receiving treatment regime $\bar{a}$, given the full data $X = (Y_{\bar{a}}, L_{\bar{a}} : \bar{a} \in \mathcal{A})$.

By the curse of dimensionality it will not be possible (even when $g(\bar{a} \mid X)$ would be known) to estimate treatment specific distributions of $Y_{\bar{a}}$ nonparametrically. Therefore we will need to assume a MSM such as:

$$E(Y_{\bar{a}} \mid V) = m_{\beta}(V, \text{sum}(\bar{a})),$$

(e.g. $\beta_0 + \beta_1 \text{sum}(\bar{a})$) where $sum(\bar{a})$ is some summary measure of $\bar{a}$ which is believed to have an effect on the conditional mean of $Y_{\bar{a}}$, within strata of $V$.

For example, if $a(k)$ is the dose of a particular treatment received at time $k$, then $\text{sum}(\bar{a}) = \sum_{k=0}^{K} a_k$ is the cumulative dose through end of follow up for a subject receiving treatment regime $\bar{a}$.

The causal parameter $\beta$ is of important policy interest: e.g $Y = 1$ when subject has detectable HIV-serum in blood at end of follow up and $a(j) = 1$ if the subject received AZT at time $j$.

# IPTW-ESTIMATOR IN MSM MODEL FOR ONE SINGLE OUTCOME

Consider the regression model:

$$E(Y \mid \bar{A}, V) = m_\beta(V, \text{sum}(\bar{A})).$$

The estimating equations for this regression model are:

$$\{h(\text{sum}(\bar{A}), V)\epsilon(\beta) : h\}.$$

The estimating equations for the corresponding MSM model $E(Y_{\bar{a}} \mid V) = m_\beta(V, \text{sum}(\bar{a}))$ are given by:

$$\left\{\frac{h(\text{sum}(\bar{A}), V)}{g(\bar{A} \mid X)}\epsilon(\beta) : h\right\}.$$

These estimating equations are unbiased if $h(\text{sum}\bar{a}, V)/g(\bar{a} \mid X) > 0$ for all $\bar{a}$.

**Remark:** If a subject's history up till point $t$ is such that certain treatments $a(t)$ have zero probability to be assigned, then one should artificially censor the subject at $t$. In this way one can artificially arrange the identifiability assumption to be true.

Define

$$\begin{aligned} SW(K) &= \frac{g(\bar{A} \mid V)}{g(\bar{A} \mid X)} \\ &= \frac{\prod_{j=0}^{K} g(A(j) \mid \bar{A}(j-1), V)}{\prod_{j=0}^{K} g(A(j) \mid \bar{A}(j-1), \bar{L}(j))}. \end{aligned}$$

In order to have a stable estimating equations which is optimal in case $V$ contains all confounders, we propose as estimating equation:

$$0 = \sum_{i=1}^{n} SW_i(K) h_{opt}(\text{sum}(\bar{A}_i), V_i) \epsilon_{\bar{A}_i}(\beta),$$

where

$$h_{opt}(\text{sum}(\bar{A}), V) = \frac{d/d\beta m_\beta(\text{sum}(\bar{A}), V)}{E(\epsilon(\beta)^2 \mid \bar{A}, V)}.$$

This estimating equation corresponds with fitting the regression model

$$E(Y \mid \bar{A}, V) = m_\beta(V, \text{sum}(\bar{A}))$$

using weights $SW_i(K)$ for subject $i$, $i = 1, \ldots, n$. We refer to these weighted estimators as IPTWE, abbreviating "Inverse Probability of Treatment Weighted Estimator".

Recall that adjusting for time-dependent confounders (thus a variable which is affected by past treatment) in the regression model will yield a biased estimate of the treatment effect: see example page 14, Robins, Hernan, Brumback (1998).

# ESTIMATION OF SUBJECT SPECIFIC WEIGHTS

Consider the case that $A(k)$ is a 1-0 variable. Then we can estimate $P(A_k = 1 \mid \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_k = \bar{l}_k)$ using a pooled logistic regression model that treats each person-day as one observation, with covariates extracted from past treatment and covariate history. This yields then an estimate of $g(\bar{A} \mid X)$. Note that this estimate is a product over time of terms $(1 - P_k)^{1-A(k)} P_k^{A(k)}$.

Similarly, one can estimate $g(\bar{A} \mid V)$ by using a pooled logistic regression model that treats each person-day as one observation, with covariate $V$ and covariates extracted from past treatment: thus not adjusting for $\bar{L}(k)$.

Give example: formula (15), (16) and (17) of Robins, Hernan, Brumback (1998).

## CENSORING BY LOSS TO FOLLOW UP

Let $C_k = 1$ if the subject was lost to follow-up by day $k$ and $C_k = 0$ otherwise. We assume that once a subject is lost to follow up, the subject does not reenter the study.

No new ideas are required to account for censoring, by viewing censoring as just another time-varying treatment and restricting the estimator above to the uncensored subjects.

The data on the uncensored subjects is now:

$$L_0, (C_0 = 0, A_0), \ldots, L_K, (C_K = 0, A_K), Y, C_{K+1} = 0.$$

Let $\bar{a}' = ((c_0, a_0), (c_1, a_1), \ldots, (c_K, a_K), c_{K+1})$ represent a treatment history: at time $j$ the subject receives joint treatment $a'_j = (c_j, a_j)$

(we define $a'_{K+1} = c_{K+1}$). As above, we define the counterfactuals $Y_{\bar{a}'}$. The only counterfactuals of interest to us are $Y_{\bar{a}'}$ for $\bar{a}'$ with $c_0 = \ldots = c_{K+1} = 0$. Therefore we only pose a MSM model for these counterfactuals. Let $Y_{\bar{a}}$ be the counterfactual $Y_{\bar{a}'}$ with real treatment components $a_j$ and $c_j = 0$, $j = 0, \ldots, K+1$. We assume the MSM model for $Y_{\bar{a}}$:

$$E(Y_{\bar{a}} \mid V) = m_\beta(V, \mathrm{sum}(\bar{a})).$$

# IPTCW-ESTIMATOR IN MSM MODEL FOR ONE SINGLE OUTCOME

Let $\Delta$ be the indicator of being uncensored: i.e. $\Delta = 1$ if and only if $C_{K+1} = 0$. The estimating equations for the MSM model $E(Y_{\overline{a}} \mid V) = m_\beta(V, \text{sum}(\bar{a}))$ are given by:

$$\left\{ \frac{h(\text{sum}(\bar{A}), V)}{g(\bar{A}' \mid X)} \epsilon(\beta) \Delta : h \right\}.$$

Define

$$\begin{aligned} SW'(K+1) &= \frac{g(\bar{A}' \mid V)}{g(\bar{A}' \mid X)} \\ &= \frac{\prod_{j=0^{K+1}} g(A'(j) \mid \bar{A}'(j-1), V)}{\prod_{j=0^{K+1}} g(A'(j) \mid \bar{A}'(j-1), \bar{L}(j))}. \end{aligned}$$

Since $a'(j) = (c(j) = 0, a(j))$ we can write

$$g(a'(j) \mid \bar{a}'(j-1), \bar{L}(j))$$

$$= g(c(j) = 0 \mid \bar{a}(j-1), \bar{c}(j-1) = 0, \bar{L}(j))$$

$$\times g(a(j) \mid \bar{a}(j-1), \bar{c}(j) = 0, \bar{L}(j))$$

and

$$g(a'(K+1) \mid \bar{a}'(K), \bar{L}(K+1))$$

$$= g(c(K+1) = 0 \mid \bar{a}(K), \bar{c}(K) = 0, \bar{L}(K+1)).$$

Therefore

$$SW'(K+1) = SW^c(K+1) SW(K),$$

where

$$SW(K) = \frac{\prod_{j=0}^{K} g(A(j) \mid \bar{A}(j-1), \bar{C}(j) = 0, V)}{\prod_{j=0}^{K+1} g(A(j) \mid \bar{A}(j-1), \bar{C}(j) = 0, \bar{L}(j))}$$

and $SW^c(K+1)$ is given by:

$$\frac{\prod_{j=0}^{K+1} g(C(j) = 0 \mid \bar{A}(j-1), \bar{C}(j-1) = 0, V)}{\prod_{j=0}^{K+1} g(C(j) = 0 \mid \bar{A}(j-1), \bar{C}(j-1) = 0, \bar{L}(j))}.$$

In order to have stable estimating equations which is optimal in case $V$ contains all confounders and nobody is censored, we propose as estimating equation:

$$0 = \sum_{i=1}^{n} SW_i'(K+1) h_{opt}(\text{sum}(\bar{A}_i), V_i)\epsilon_{\bar{A}_i}(\beta),$$

where

$$h_{opt}(\text{sum}(\bar{A}), V) = \frac{d/d\beta m_{\beta}(\text{sum}(\bar{A}), V)}{E(\epsilon(\beta)^2 \mid \bar{A}, V)}.$$

This estimating equation corresponds with fitting the regression model $E(Y \mid \bar{A}, V) = m_{\beta}(V, \text{sum}(\bar{A}$ with using weights $SW_i'(K+1)$ for subject $i$, $i = 1, \ldots, n$. We refer to these weighted estimators as "Inverse of Probability of Treatment and Censoring Weighted Estimator".

Explain that these estimators are the same as solving the estimating equation we used without censoring with $\Delta/P(\Delta = 1 \mid X, A)$.

# ESTIMATION OF SUBJECT SPECIFIC WEIGHTS

Again, we can estimate $P(C_k = 0 \mid \bar{C}_{k-1} = 0, \bar{A}(k-1), \bar{L}_k)$ and $P(C_k = 0 \mid \bar{C}_{k-1} = 0, \bar{A}(k-1), V)$ using a pooled logistic regression model that treats each person-day as one observation. Thus by fitting four logistic regression models to pooled samples one obtains an estimate of $SW'(K+1)$.

# PART V:

# MARGINAL STRUCTURAL MODELS FOR TIME-DEPENDENT TREATMENT IN SURVIVAL ANALYSIS

# DATA

Let $A(j)$ be treatment the subject received at time $j$. Let $L(j)$ be time-dependent covariates collected on the subject at time $j$, where $L(j)$ occurs right before the treatment assignment $A(j)$. Let the outcome of interest be the survival time $T$ of the subject. A particular application one can keep in mind is a longitudinal study in which a HIV-infected subject is followed up till death $T$, $A(t)$ is a dichotomous variable indicating whether a patient is on prophylaxis treatment at day $t$, $L(t)$ is a vector of measured risk factors for survival such as CD4 count, white blodd cell count and number of Pneumonia (PCP) bouts.

The observed data on a subject is thus:

$$(T, \bar{A}(T), \bar{L}(T)).$$

We are concerned with estimation of causal effects of $\bar{A}$ on survival $T$. A useful alternative way of representing this data structure is to define $Y(j)$ as the indicator of failure at time $j$ and define the data as:

$$(\bar{A}(T), \bar{Y}(T), \bar{L}(T)).$$

Let $V$ be a subset of the baseline covariates $L(0)$. Let $T_{\bar{a}}$ be the counterfactual value of $T$ that would have been observed had the subject received treatment history $\bar{a} = (a_0, \ldots, a_K)$. We have $T_{\bar{a}} = T_{\bar{a}(T),0}$. We can also define counterfactuals $L_{\bar{a}}$ which denotes the process $L$ that would have been observed if the subject had received treatment $\bar{a}$. Again, $L_{\bar{a}}(t) = L_{\bar{a}(t),0}(t)$. The $Y_{\bar{a}}$, $\bar{a} \in \mathcal{A}$, are the counterfactuals of interest.

We will assume that treatment is sequentially randomized: for each possible treatment regime $\bar{a}$ (consistent with the observed history)

$$A(k) \perp Y_{\bar{a}} \mid \bar{A}(k-1), \bar{L}(k).$$

In other words, for each $k$

$$g(a(k) \mid (Y_{\bar{a}} : \bar{a} \in \mathcal{A}), \bar{A}(k-1), \bar{L}(k))$$

$$= g(A(k) \mid \bar{A}(k-1), \bar{L}(k)).$$

We define:

$$g(\bar{a} \mid X) = \prod_{k=0}^{K} g(a(k) \mid \bar{A}(k-1), \bar{L}(k)),$$

which one can think of as the conditional probability on receiving treatment regime $\bar{a}$, given the full data $X = (Y_{\bar{a}}, L_{\bar{a}} : \bar{a} \in \mathcal{A})$.

## MARGINAL STRUCTURAL COX MODEL

In the absence of time-dependent confounding one could use a time-dependent Cox-proportional hazards model:

$$\lambda_T(t \mid \bar{A}(t), V) = \lambda_0(t)\, \text{exp}(\gamma_1 A(t) + \gamma_2^\top V).$$

Here $\lambda_T(t \mid \bar{A}(t), V)$ is the hazard of death at time $t$ from start of follow up conditional on treatment history $\bar{A}(t)$ and pretreatment covariates $V$, and $\lambda_0(t)$ is an unspecified baseline hazard function. For example, $V$ could include the log of baseline CD4-count, log of baseline white blood count.

In the absence of time-dependent confounding one can then estimate $\gamma$ with the solution of the partial likelihood score equation for $\gamma$. Since the partial likelihood is a product over time from $t = 0$ till $\infty$ the score equation is an sum over time. So let's represent the score equation (for one subject) as $\sum_t U(\bar{A}(t), \bar{Y}(t), V \mid \gamma)$.

The corresponding marginal structural Cox-proportional hazards model is given by:

$$\lambda_{T_{\bar{a}}}(t \mid V) = \lambda_0(t) \, \mathrm{exp}(\beta_1 a(t) + \beta_2 V),$$

where $\lambda_{T_{\bar{a}}}(t \mid V)$ is the hazard of death at $t$ among subjects with pretreatment covariates $V$ had, contrary to the fact, all subjects followed treatment regime $\bar{a}$.

Define

$$\begin{aligned} SW(t) &= \frac{g(\bar{A}(t) \mid V)}{g(\bar{A}(t) \mid X)} \\ &= \frac{\prod_{j=0}^{t} g(A(j) \mid \bar{A}(j-1), V)}{\prod_{j=0}^{K} g(A(j) \mid \bar{A}(j-1), \bar{L}(j))}. \end{aligned}$$

In order to have a stable estimating equation which is optimal in case we do not have time-dependent confounding, we propose as estimating equation:

$$0 = \sum_{i=1}^{n} \sum_{t} SW_i(t) U(\bar{A}_i(t), \bar{Y}_i(t), V_i \mid \gamma).$$

This corresponds with fitting the time-dependent Cox model with each subjects data line $(A_i(t), Y_i(t), L$ weighted with $W_i(t) = SW_i(t)$ with $t$ running from 0 till $T_i$.

# MARGINAL STRUCTURAL lOGISTIC REGRESSION MODEL

If time is discrete, i.e. many subjects die at the same time, then the Cox-model is not appropriate, but one should use a discrete survival time model.

In this case one could model the discrete hazard with a logistic regression model:

$$\text{logit}(P(Y(t) = 1 \mid Y(t-1) = 0, \bar{A}(t-1), V)$$

$$= \beta_0(t) + \beta_1 A(t-1) + \beta_2 V,$$

where $\beta_0(t)$ is an unspecified baseline function. If the time unit becomes finer and finer, then this model approximates the Cox proportional hazards model with $\exp(\beta_0(t))$ representing the cumulative baseline hazard.

This model can be fit with pooled logistic regression treating each person day as an observation: this also provides the correct confidence intervals.

The corresponding marginal structural model is given by:

$$\text{logit}(P(Y_{\bar{a}}(t) = 1 \mid Y_{\bar{a}}(t-1) = 0, V)$$

$$= \beta_0(t) + \beta_1 a(t-1) + \beta_2 V.$$

The causal parameters $\beta$ can be fit with weighted pooled logistic regression treating each person day $t$ as an observation with weights $SW(t)$. To obtain conservative confidence intervals one needs to view the data as repeated measures and therefore one should fit the model with a generalized estimating equations program (e.g. option 'repeated' in SAS Proc Genmod).

# CENSORING BY LOSS TO FOLLOW UP

Let $C_k = 1$ if the subject was lost to follow-up by day $k$ and $C_k = 0$ otherwise. We assume that once a subject is lost to follow up, the subject does not reenter the study.

No new ideas are required to account for censoring, by viewing censoring as just another time-varying treatment and restricting the estimator above to the uncensored subjects. At time $j$ the subject receives joint treatment $a'_j = (c_j, a_j)$.

As above, we define the counterfactuals $Y_{\overline{a}'}$. The only counterfactuals of interest to us are $Y_{\overline{a}'}$ for $\overline{a}'$ with $c_0 = \ldots = c_{K+1} = 0$. Therefore we only pose the logistic regression or Cox-proportional hazards MSM model for these counterfactuals.

Let $\Delta$ be the indicator of being uncensored: i.e. $\Delta = 1$ if and only if $C_{K+1} = 0$. Define

$$\begin{aligned} SW'(t) &= \frac{g(\bar{A}'(t) \mid V)}{g(\bar{A}'(t) \mid X)} \\ &= \frac{\prod_{j=0}^{t} g(A'(j) \mid \bar{A}'(j-1), V)}{\prod_{j=0}^{t} g(A'(j) \mid \bar{A}'(j-1), \bar{L}(j))}. \end{aligned}$$

Since $a'(j) = (c(j) = 0, a(j))$ we can write

$$g(a'(j) \mid \bar{a}'(j-1), \bar{L}(j))$$

$$= g(c(j) = 0 \mid \bar{a}(j-1), \bar{c}(j-1) = 0, \bar{L}(j))$$

$$\times g(a(j) \mid \bar{a}(j-1), \bar{c}(j) = 0, \bar{L}(j)).$$

Therefore

$$SW'(t) = SW^c(t) SW(t),$$

where

$$SW(t) = \frac{\prod_{j=0}^{t} g(A(j) \mid \bar{A}(j-1), \bar{C}(j) = 0, V)}{\prod_{j=0}^{t} g(A(j) \mid \bar{A}(j-1), \bar{C}(j) = 0, \bar{L}(j))}$$

and $SW^c(t)$ is given by:

$$\frac{\prod_{j=0}^{t} g(C(j)=0 \mid \bar{A}(j-1), \bar{C}(j-1)=0, V)}{\prod_{j=0}^{t} g(C(j)=0 \mid \bar{A}(j-1), \bar{C}(j-1)=0, \bar{L}(j))}.$$

One estimates $\beta$ with weighted pooled logistic regression treating each person day t as an observation with weights $\Delta SW'(t)$.

# INSTRUMENTAL VARIABLES IN REGRESSION

Suppose that $Y = m(X \mid \beta) + \epsilon$, where $E\epsilon = 0$ but $E(\epsilon \mid X) \neq 0$. For example, $X$ might be the actual treatment taken by the subject, $Y$ is the outcome of interest and $X$ might be based on unobserved variables related to the error. Then the standard (naive) estimating equation $h(X)\epsilon(\beta)$ might result in a biased estimator. Let $Z$ be a variable satisfying $E(\epsilon(\beta) \mid Z) = E(\epsilon(\beta))$; for example, $Z$ is independent of $\epsilon(\beta)$. In our example, one could think of $Z$ being a randomly assigned treatment arm. Then one can use as estimating equation

$$g(Z)\epsilon(\beta). \tag{2}$$

If the matrix $E(g(Z)\frac{d}{d\beta}\epsilon(\beta))$ is invertible, then under standard regularity conditions, the corresponding estimator is asymptotically linear with influence curve

$$\{Eg(Z)d/d\beta m(X \mid \beta)\}^{-1}g(Z)\epsilon(\beta).$$

This invertibility condition requires that $E(g(Z)d/d\beta m(X \mid \beta)) \neq Eg(Z)Ed/d\beta m(X \mid \beta)$. In other words, this estimating equation can only be informative if $Z$ is related to $X$. The random variable $Z$ is often referred to as an instrumental variable. Thus in regression problems where one expects dependence between the residual and $X$ one can salvage estimation by finding a variable $Z$ which is unrelated to the residual but related to $X$.

# CAUSAL INFERENCE WITH NON-COMPLIANCE IN POINT TREATMENT STUDIES

Let $R$ be the treatment assigned to the subject and we assume that $R$ is completely randomized. Let $A$ be the treatment the subject actually uses. Let $Y$ be the outcome of interest and suppose that we also observe some covariates $W$. Thus the observed data is $(Y, R, A, W)$. By non-compliance $A$ can be different from $R$ and $A$ can be confounded by unmeasured confounders. Let $X = ((Y_a : a), W)$ be the treatment specific counterfactual outcomes and the covariate vector.

Consider the marginal structural model

$$Y_a = \beta_0 + \beta_1 a + \epsilon, \text{ where } E(\epsilon) = 0.$$

Note that $\epsilon = Y_0 - \beta_0$ so that $\beta_0 = EY_0$. This marginal structural model is equivalent with

$E(Y_a - Y_0) = \beta_1 a.$

It also corresponds with the following observed data regression model

$$Y = Y_A = \beta_0 + \beta_1 A + \epsilon, \text{ where } E\epsilon(\beta) = 0.$$

Thus estimation of $\beta_1, \beta_2$ corresponds with linear regression of $Y$ on $A$ but with an error term which depends on $A$ since the actual selected treatment $A$ might have been based on $Y_0$.

This suggests to use the instrumental variable method to estimate $(\beta_1, \beta_2)$ using $R$ as instrumental variable. Notice that indeed $R$ is independent of $\epsilon$ and (strongly) related to $A$. Thus our estimating equations are of the type: for any given $\phi$

$$\phi(R)\{Y - \beta_0 - \beta_1 A\}.$$

The unbiasedness of this estimating equation follows from the fact that at the true $\beta$ $R$ is independent of $\epsilon(\beta) = Y - \beta_0 - \beta_1 A$ and that $E\epsilon(\beta) = 0$. Alternatively, we could use as estimating equation:

$$\{\phi(R) - E\phi(R)\}\{Y - \beta_1 A\}.$$

If $R$ has only two outcomes $0, 1$, then there exists only one estimating equation (i.e. $\phi$) and therefore one can only identify $\beta_1$. In general, the dimension of our causal model parameter $\beta$ needs to be restricted by the actual number of estimating equations we can come up with. If $R$ has $k$ possible outcomes, then we can come up $k-1$ choices of $\phi$ . If covariates are available, then we have $k-1$ estimating equations for each strata identified by e.g. $V = v$. By assuming that the causal model does not heavily depend on the strate $V = v$, e.g. $E(Y_a - Y_0 \mid V) = \beta_1 a + \beta_2 V$, this approach makes it possible to model the effect of $a$ more flexible.

# CAUSAL EFFECT AMONG COMPLIERS

Assume the following model:

$$E(Y_a - Y_0 \mid R, A = a) = \beta_0 a + \beta_1 R.$$

Note that the unknown parameter $\beta = (\beta_0, \beta_1)$ defines, in particular, the causal effect of treatment $A$ among the compliers. Let $Y_0(\beta) = Y - \beta_0 A - \beta_1 R$ which represents the outcome $Y$ blipped down to $Y_0$. The instrumental variable method suggests the following estimating equation for $\beta$:

$$(\phi(R) - E\phi(R))Y_0(\beta). \qquad (3)$$

Since $E(Y_0(\beta) \mid R, A = a) = E(Y_0 \mid R, A = a)$ it follows that

$$\begin{aligned} E\{(\phi(R) - E\phi(R))Y_0(\beta)\} &= E\{(\phi(R) - E\phi(R))Y_0\} \\ &= = 0 \end{aligned}$$

since $E(\phi(R) \mid Y_0) = E\phi(R)$.

(We used that $Y_{RA} = Y_A$)

# CAUSAL INFERENCE WITH NON-COMPLIANCE IN LONGITUDINAL STUDIES

**Data:** On each subject we collect the following data over time

$$R, L_0, A_0, L_1, A_1, \ldots, L_K, A_K, Y,$$

where $(L_j, A_j)$ represents covariates and treatment at time $j$, $j = 0, \ldots, K$, and $R = A_{-1}$ is the randomly assigned treatment arm. We model the so called blip function conditional on the past:

$$E(Y_{\bar{A}_j,0} - Y_{\bar{A}_{j-1},0} \mid \bar{A}_j = \bar{a}_j, \bar{L}_j = \bar{l}_j) = \beta_j(\bar{a}_j, \bar{l}_j \mid \beta),$$

where the blip function $\beta_j$ is parametrized by a finite dimensional parameter vector $\beta$ which is common to each $\beta_j$, $j = 1, \ldots, K$. In words, this blip function is the expected value of the difference of two counterfactuals only differing by one blip in their treatment, given the observed past of the subject.

Consider now the completely blipped down version of $Y$

$$Y_0(\beta) = Y_{\bar{A}_K} - \sum_{l=1}^{K} \beta_l(\bar{A}_l, \bar{L}_l).$$

At the true $\beta$ $E(Y_0(\beta) \mid R) = EY_0$.

Using $R$ as an instrumental variable suggests the following estimating equations: for any given $\phi$

$$\{\phi(R) - E\phi(R)\}\{Y_0(\beta)\}.$$

If $R$ has only two outcomes $0, 1$, then there exists only one estimating equation (i.e. $\phi$) and therefore one can only identify $\beta_1$. In general, the dimension of our causal model parameter $\beta$ needs to be restricted by the actual number of estimating equations we can come up with. If $R$ has $k$ possible outcomes, then we can come up $k-1$ choices of $\phi$ . If covariates are available, then we have $k-1$ estimating equations for each strata identified by e.g. $V = v$. By assuming that the causal model does not heavily depend on the strate $V = v$, e.g. $E(Y_a - Y_0 \mid V) = \beta_1 a + \beta_2 V$, this approach makes it possible to model the effect of $a$ more flexible.

# STRUCTURAL NESTED MEAN MODELS IN LONGITUDINAL STUDIES

**Data:** On each subject we collect the following data over time

$$L_0, A_0, L_1, A_1, \ldots, L_K, A_K, Y,$$

where $(L_j, A_j)$ represents covariates and treatment, respectively, at time $j$, $j = 0, \ldots, K$, and $Y$ is the outcome of interest. Let $\bar{A}_j = (A_0, \ldots, A_j)$ and $\bar{L}_j = (L_0, \ldots, L_j)$, $j = 0, \ldots, K$.

For each possible treatment regime $\bar{a} = (a_0, \ldots, a_K)$ we define $(Y_{\bar{a}}, \bar{L}_{\bar{a}}$ as the counterfactual outcome of $(Y, \bar{L})$ if, possibly contrary to the fact, the subject would have received treatment regime $\bar{a}$. Thus $(Y, \bar{L}) = (Y_{\bar{A}}, \bar{L}_{\bar{A}})$.

We assume the following model. Firstly, we assume the sequential randomization assumption which states that $A_j \perp \{Y_{\bar{a}}, \bar{L}_{\bar{a}} : \bar{a}\}$, given the observed past $\bar{A}_{j-1}, \bar{L}_j$, where $\bar{a}$ ranges over treatment regimes with $\bar{a}_{j-1} = \bar{A}_{j-1}$. In addition, we model the so called *blip function* conditional on the past:

$$E(Y_{\bar{A}_j,0} - Y_{\bar{A}_{j-1},0} \mid \bar{A}_j = \bar{a}_j, \bar{L}_j = \bar{l}_j) = \beta_j(\bar{a}_j, \bar{l}_j \mid \beta),$$

where the blip function $\beta_j$ is parametrized by a finite dimensional parameter vector $\beta$ which is common to each $\beta_j$, $j = 1, \ldots, K$. In words, this blip function is the expected value of the difference of two counterfactuals only differing by one blip in their treatment, given the observed past.

The idea above of using an instrumental variable to obtain an unbiased estimating equation can be generalized to construct unbiased estimating equations of the blip function in structural nested mean models. We view the total data generating experiment as a sequential experiment over time, where at time $j$ one conditions on the observed past $\bar{A}(j-1), \bar{L}(j)$. Experiment $j$ corresponds with drawing the data after $A_{j-1}$ and ending with generating $A_j$, where we know that $A_j$ is assigned completely at random, given the past. For each $j$ one constructs a residual which has mean zero conditonal on the past and is unrelated to $A_j$ which will play the role of the instrumental variable.

Consider the blipped down version of $Y$

$$Y_{j-1}(\beta) = Y_{\bar{A}_K} - \sum_{l=j}^{K} \beta_l(\bar{A}_l, \bar{L}_l).$$

Define the residual:

$$\epsilon_{j-1}(\beta) \equiv Y_{j-1}(\beta) - E(Y_{j-1}(\beta) \mid \bar{A}_{j-1}, \bar{L}_j),$$

which has expectation zero, given $\bar{A}_{j-1}, \bar{L}_j$.

Notice that $A_j$ is related to the covariates $\sum_{l=j}^{K} \beta_l(\bar{A}_l, \bar{L}_l)$ and using 1) that $Y_{j-1}(\beta)$ represents the counterfactual $Y_{\bar{A}_{j-1},0}$ and 2) the sequential randomization assumption we will be able to show that

$$E(\epsilon_{j-1}(\beta) \mid \bar{A}_{j-1}, A_j, \bar{L}_j) = E(\epsilon_{j-1}(\beta) \mid \bar{A}_{j-1}, \bar{L}_j). \tag{4}$$

Thus $A_j$ is unrelated (in the expectation sense) to the residual, given the observed past. This proves that we can use $A_j$ as instrumental variable and thus use as estimating equation: for each function $g$

$$\epsilon_{j-1}(\beta) g(\bar{A}_j, \bar{L}_j) = 0, \ j = 1, \ldots, K.$$

To see that the estimating equation is unbiased just condition on $\bar{A}_j, \bar{L}_j$ and use that $E(Y_{j-1}(\beta) \mid \bar{A}_j, \bar{L}_j) = E(Y_{j-1}(\beta) \mid \bar{A}_{j-1}, \bar{L}_j)$.

A natural way of combining these $K$ instrumental estimating equations corresponding with experiment $j = 1, \ldots, K$ to one estimating equation for $\beta$ is to use as estimating equation

$$\sum_{j=1}^{K} \epsilon_{j-1}(\beta) g_j(\bar{A}_j, \bar{L}_j) = 0.$$

We can extend this class of estimating equations as follows:

$$\sum_{j=1}^{K} \left\{ Y_{j-1}(\beta) - \Phi(\bar{A}_{j-1}, \bar{L}_j) \right\}$$
$$\left\{ g(\bar{A}_j, \bar{L}_j) - E(g(\bar{A}_j, \bar{L}_j) \mid \bar{A}_{j-1}, \bar{L}_j) \right\},$$

where $\phi$ and $g$ are user supplied.

We will now show that indeed

$$E(Y_{j-1}(\beta) \mid \bar{A}_{j-1}, \bar{L}_j, A_j) = E(Y_{j-1}(\beta) \mid \bar{A}_{j-1}, \bar{L}_j).$$

We have

$$E\left(Y_{\bar{A}} - Y_{\bar{A}_{j-1},0} - \beta \sum_{l=j}^{k} A_l \mid \bar{A}_j, \bar{L}_j\right)$$

$$= \textstyle\sum_{m=0}^{k-j} E(Y_{\bar{A}_{m+j},0} - Y_{\bar{A}_{m+j-1},0} - \beta_1 A_{m+j} \mid \bar{A}_j, \bar{L}_j)$$

$$= \sum_{m=0}^{k-j} E\left\{E\left(Y_{\bar{A}_{m+j},0} - Y_{\bar{A}_{m+j-1},0} - \beta_1 A_{m+j} \mid \right.\right.$$

$$\left.\left.\bar{A}_{m+j}, \bar{L}_{m+j}\right) \mid \bar{A}_j, \bar{L}_j\right\}$$

$$= 0.$$

Thus

$$E(Y_{j-1}(\beta) \mid \bar{A}_j, \bar{L}_j) = E(Y_{\bar{A}_{j-1},0} \mid \bar{A}_j, \bar{L}_j).$$

By the sequential randomization assumption the latter equals $E(Y_{\bar{A}_{j-1},0} \mid \bar{A}_{j-1}, \bar{L}_j)$.

# ESTIMATING COUNTERFACTUAL EXPECTATIONS

Above we provided an estimating equation for the blip function parameter $\beta$. Suppose now that we are concerned with estimating $E(Y_{\bar{a}})$ for a given treatment regime $\bar{a} = (a_0, \ldots, a_K)$. In order to derive an estimator of this parameter we will do as if $\beta$, i.e. the set of blip functions $\beta_j$, is known. The actual proposed estimator of $E(Y_{\bar{a}})$ is obtained by substituting an estimate for $\beta$.

For each subject construct the following variable

$$Y_0(\beta) = Y_{\bar{A}} - \sum_{l=1}^{K} \beta_l(\bar{A}_l, \bar{L}_l).$$

The variable $Y_0(\beta)$ represents a substitute for the variable $Y_0$ one would have seen if the subject had never been treated. As above one can show that $EY_0(\beta) = EY_0$, where $Y_0$ is the counterfactual outcome $Y$ under regime $\bar{a} = 0$.

Consider now the random variable:

$$Y_{\bar{a}}(\beta) \equiv Y_0(\beta) + \sum_{l=1}^{K} \beta_l(\bar{a}_l, \bar{L}_{l,\bar{a}}),$$

where $\bar{L}_{l,\bar{a}}$ is the counterfactual of $(L_1, \ldots, L_l)$ corresponding with treatment regime $(a_0, \ldots, a_l)$. Note that this variable is random through $Y_0(\beta)$ and $\bar{L}_{K,\bar{a}}$, but $\bar{A}_l$ is fixed at $\bar{a}_l$. We note that:

$$E\left\{Y_0(\beta) + \sum_{l=1}^{K} \beta_l(\bar{a}_l, \bar{L}_{l,\bar{a}})\right\}$$

$$= EY_0 + \sum_{l=1}^{K} E_{\bar{L}_{l,\bar{a}}} E(Y_{\bar{a}_l,0} - Y_{\bar{a}_{l-1},0} \mid \bar{A}_l = \bar{a}_l, \bar{L}_{l,\bar{a}})$$

$$= EY_0 + EY_{\bar{a}} - EY_0 = EY_{\bar{a}}.$$

Thus the random variable $Y_{\bar{a}}(\beta)$ has the same expectation as the treatment specific counterfactual $Y_{\bar{a}}$. Thus it remains to understand how to estimate $EY_{\bar{a}}(\beta)$.

Note that the expectation of $\beta_l(\bar{A}_l, \bar{L}_l)$ is taken in the world where everybody get assigned treatment $\bar{A} = \bar{a}$, which comes down to integrating w.r.t. the joint distribution of the counterfactuals of $L_0, L_{1a_1}, L_{2a_1a_2}, \ldots, L_{l,\bar{a}_l}$ and setting $\bar{A}_K = \bar{a}_K$. This joint distribution is obtained with the general $G$-computation formula which we will give now.

First write down the density representation for $L_0, A_0, L_1, A_1, \ldots, L_l, A_l$:

$$f(L_0)f(A_0 \mid L_0)f(L_1 \mid \bar{A}_0, L_0)f(A_1 \mid \bar{A}_0, \bar{L}_1)$$

$$\ldots f(L_l \mid \bar{L}_{l-1}, \bar{A}_{l-1})f(A_l \mid \bar{L}_l, \bar{A}_{l-1}).$$

Replacing $f(A_j \mid \bar{A}_{j-1}, \bar{L}_j)$ by a degenerate distribution at $A_j = a_j$, $j = 0, \ldots, l$, results in the wished joint density $P(L_0 = s_0, L_{1a_1} = s_1, L_{2a_1a_2} = s_2, \ldots, L_{l,\bar{a}_l} = s_l)$ given by:

$$\prod_{j=0}^{l} P(L_j = s_j \mid \bar{A}_{j-1} = \bar{a}_{j-1}, \bar{L}_{j-1} = s_{j-1}).$$

The latter formula is referred to as the G-computation formula and indeed equals the counterfactual density under the sequential randomization assumption.

We conclude that we have the following formula for $EY_{\bar{a}}$:

$$EY_{\bar{a}} = EY_0(\beta) + \sum_{l=1}^{K} \int_{s_1,\ldots,s_l} \beta_l(\bar{a}_l, \bar{s}_l)$$
$$\prod_{j=0}^{l} P(L_j = s_j \mid \bar{A}_{j-1} = \bar{a}_{j-1}, \bar{L}_{j-1} = s_{j-1})$$

This formula expresses the counterfactual expectation $EY_{\bar{a}}$ in terms of observed data distributions and the blip function. Consequently, we can use this formula to estimate $EY_{\bar{a}}$. Beyond estimation of the blip function it requires estimation of the conditional distribution of $L_j$, given the past.

For testing the presence of a treatment effect one is only concerned with estimation of the blip function itself which does not require modelling of covariate distributions. If one uses the formula to estimate $EY_{\bar{a}}$ for various $\bar{a}$, then these estimates are protected agains misspecification of the covariate distributions under the null-hypothesis of no-treatment effect.

# EXTENSION TO DYNAMIC REGIMES

For a given set of rules $\bar{d} = (d_1(\cdot), \ldots, d_K(\cdot))$ let $Y_{\bar{d}}$ be the counterfactual outcome of $Y$ if one follows the rules $A_j = d_j(\bar{A}_{j-1}, \bar{L}_j)$. Suppose we want to estimate $EY_{\bar{d}}$.

We already provided estimators for the blip function and we can also still define $Y_0(\beta)$ as above. Define

$$Y_{\bar{d}}(\beta) \equiv Y_0(\beta) + \sum_{l=0}^{K} \beta_l(\bar{A}_l, \bar{L}_{l,\bar{d}_l}),$$

where $(\bar{A}_l, \bar{L}_{l,\bar{d}_l})$ follows the counterfactual distribution one would observe in the hypothetical world where everybody follows the dynamic treatment regime $\bar{d}$. As above one can show that the expectation of $Y_{\bar{d}}(\beta)$ equals the expectation of $Y_{\bar{d}}$. Thus it remains to estimate $EY_{\bar{d}}(\beta)$.

This counterfactual distribution of $(\bar{A}_l, \bar{L}_{l,\bar{d}_l})$ is obtained with the general $G$-computation formula. First write down the density representation for the data $L_0, A_0, L_1, A_1, \ldots, L_l, A_l$:

$$f(L_0)f(A_0 \mid L_0)f(L_1 \mid \bar{A}_0, L_0)f(A_1 \mid \bar{A}_0, \bar{L}_1) \ldots f(L_l \mid$$

Replacing $f(A_j \mid \bar{A}_{j-1} = \bar{a}_{j-1}, \bar{L}_j = \bar{s}_j)$ by a degenerate distribution at $d_j(\bar{a}_{j-1}, \bar{s}_j)$, $j = 0, \ldots, l$, results in the wished joint density $P(L_0 = s_0, L_{1d_1} = s_1, L_{2d_1d_2} = s_2, \ldots, L_{l,\bar{d}_l} = s_l)$ given by:

$$\prod_{j=0}^{l} P(L_j = s_j \mid \bar{A}_{j-1} = d_j(\bar{a}_{j-1}, \bar{L}_j), \bar{L}_{j-1} = \bar{s}_{j-1}).$$

The latter formula is referred to as the G-computation formula and indeed equals the counterfactual density under the sequential randomization assumption.

We conclude that we have the following formula for $EY_{\bar{d}}$:

$$EY_{\bar{d}} = EY_0(\beta) + \sum_{l=1}^{K} \int_{s_1,\ldots,s_l} \beta_l(\bar{a}_l, \bar{s}_l)$$
$$\prod_{j=0}^{l} P(L_j = s_j \mid \bar{A}_{j-1} = d_j(\bar{a}_{j-1}, \bar{s}_j), \bar{L}_{j-1} = \bar{s}_{j-1})$$

Given estimates of the conditional distributions of $L_j$, given the past, for $j = 0, \ldots, K$, given $\beta$ and thus $Y_0(\beta)$ one can evaluate this multivariate integral by simply simulating a large number of the variables $Y_{\bar{d}}(\beta)$. This avoids the need of numerical integration.