Structural Nested Models in Reproductive Epidemiology

Ashley I Naimi, PhD
ashley.naimi@pitt.edu

June 20 2016

Acknowledgements

- Steve Cole (UNC)
- Robert Platt (McGill)
- Miguel Hernán (HSPH)
- Stijn Vansteelandt (U Ghent)
- Erica Moodie (McGill)

- Ya Hui Yu & Hsinghua Lin (Pitt)

All errors, oversights,
and obscurities are my own.

Outline

1 Complications of Time & Causal Inference

2 Mediation Analysis

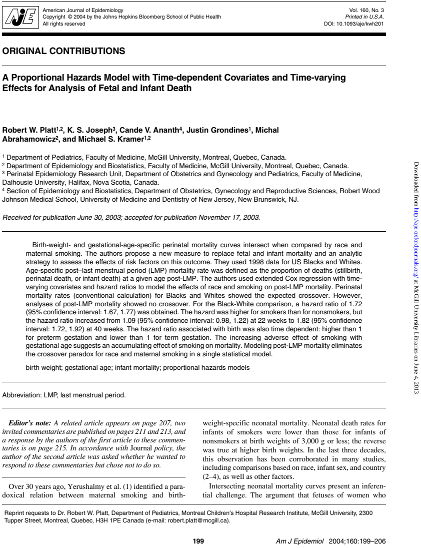
3 G Methods

4 G Estimation of Structural Nested Models

   Working Example 1 (Linear SNMM)

   Working Example 2 (Log-Linear SNMM & Mediation)

Timing is Everything!



American Journal of Epidemiology
Copyright © 2004 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved

Vol. 160, No. 3
Printed in U.S.A.
DOI: 10.1093/aje/kwh201

ORIGINAL CONTRIBUTIONS

A Proportional Hazards Model with Time-dependent Covariates and Time-varying Effects for Analysis of Fetal and Infant Death

Robert W. Platt[1,2], K. S. Joseph[3], Cande V. Ananth[4], Justin Grondines[1], Michal Abrahamowicz[2], and Michael S. Kramer[1,2]

[1] Department of Pediatrics, Faculty of Medicine, McGill University, Montreal, Quebec, Canada.
[2] Department of Epidemiology and Biostatistics, Faculty of Medicine, McGill University, Montreal, Quebec, Canada.
[3] Perinatal Epidemiology Research Unit, Department of Obstetrics and Gynecology and Pediatrics, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada.
[4] Section of Epidemiology and Biostatistics, Department of Obstetrics, Gynecology and Reproductive Sciences, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, New Brunswick, NJ.

Received for publication June 30, 2003; accepted for publication November 17, 2003.

Birth-weight- and gestational-age-specific perinatal mortality curves intersect when compared by race and maternal smoking. The authors propose a new measure to replace fetal and infant mortality and an analytic strategy to assess the effects of risk factors on this outcome. They used 1998 data for US Blacks and Whites. Age-specific post–last menstrual period (LMP) mortality rate was defined as the proportion of deaths (stillbirth, perinatal death, or infant death) at a given age post-LMP. The authors used extended Cox regression with time-varying covariates and hazard ratios to model the effects of race and smoking on post-LMP mortality. Perinatal mortality rates (conventional calculation) for Blacks and Whites showed the expected crossover. However, analyses of post-LMP mortality showed no crossover. For the Black-White comparison, a hazard ratio of 1.72 (95% confidence interval: 1.67, 1.77) was obtained. The hazard was higher for smokers than for nonsmokers, but the hazard ratio increased from 1.09 (95% confidence interval: 0.98, 1.22) at 22 weeks to 1.82 (95% confidence interval: 1.72, 1.92) at 40 weeks. The hazard ratio associated with birth was also time dependent: higher than 1 for preterm gestation and lower than 1 for term gestation. The increasing adverse effect of smoking with gestational age suggests an accumulating effect of smoking on mortality. Modeling post-LMP mortality eliminates the crossover paradox for race and maternal smoking in a single statistical model.

birth weight; gestational age; infant mortality; proportional hazards models

Abbreviation: LMP, last menstrual period.

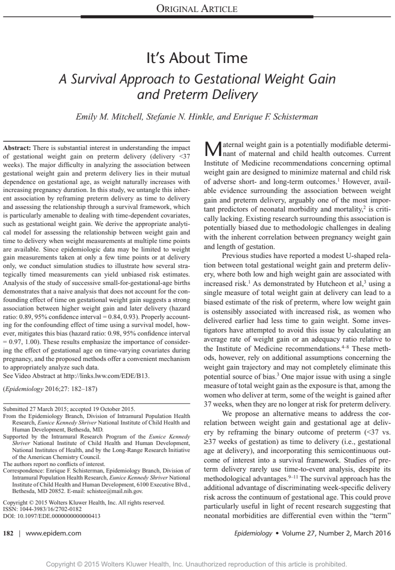
***Editor's note:** A related article appears on page 207, two invited commentaries are published on pages 211 and 213, and a response by the authors of the first article to these commentaries is on page 215. In accordance with* Journal *policy, the author of the second article was asked whether he wanted to respond to these commentaries but chose not to do so.*

Over 30 years ago, Yerushalmy et al. (1) identified a paradoxical relation between maternal smoking and birth-weight-specific neonatal mortality. Neonatal death rates for infants of smokers were lower than those for infants of nonsmokers at birth weights of 3,000 g or less; the reverse was true at higher birth weights. In the last three decades, this observation has been corroborated in many studies, including comparisons based on race, infant sex, and country (2–4), as well as other factors.

Intersecting neonatal mortality curves present an inferential challenge. The argument that fetuses of women who

Reprint requests to Dr. Robert W. Platt, Department of Pediatrics, Montreal Children's Hospital Research Institute, McGill University, 2300 Tupper Street, Montreal, Quebec, H3H 1PE Canada (e-mail: robert.platt@mcgill.ca).

199 *Am J Epidemiol* 2004;160:199–206



ORIGINAL ARTICLE

It's About Time

A Survival Approach to Gestational Weight Gain and Preterm Delivery

Emily M. Mitchell, Stefanie N. Hinkle, and Enrique F. Schisterman

Abstract: There is substantial interest in understanding the impact of gestational weight gain on preterm delivery (delivery <37 weeks). The major difficulty in analyzing the association between gestational weight gain and preterm delivery lies in their mutual dependence on gestational age, as weight naturally increases with increasing pregnancy duration. In this study, we untangle this inherent association by reframing preterm delivery as time to delivery and assessing the relationship through a survival framework, which is particularly amenable to dealing with time-dependent covariates, such as gestational weight gain. We derive the appropriate analytical model for assessing the relationship between weight gain and time to delivery when weight measurements at multiple time points are available. Since epidemiologic data may be limited to weight gain measurements taken at only a few time points or at delivery only, we conduct simulation studies to illustrate how several strategically timed measurements can yield unbiased risk estimates. Analysis of the study of successive small-for-gestational-age births demonstrates that a naive analysis that does not account for the confounding effect of time on gestational weight gain suggests a strong association between higher weight gain and later delivery (hazard ratio: 0.89, 95% confidence interval = 0.84, 0.93). Properly accounting for the confounding effect of time using a survival model, however, mitigates this bias (hazard ratio: 0.98, 95% confidence interval = 0.97, 1.00). These results emphasize the importance of considering the effect of gestational age on time-varying covariates during pregnancy, and the proposed methods offer a convenient mechanism to appropriately analyze such data.
See Video Abstract at http://links.lww.com/EDE/B13.

(*Epidemiology* 2016;27: 182–187)

Maternal weight gain is a potentially modifiable determinant of maternal and child health outcomes. Current Institute of Medicine recommendations concerning optimal weight gain are designed to minimize maternal and child risk of adverse short- and long-term outcomes.[1] However, available evidence surrounding the association between weight gain and preterm delivery, arguably one of the most important predictors of neonatal morbidity and mortality,[2] is critically lacking. Existing research surrounding this association is potentially biased due to methodologic challenges in dealing with the inherent correlation between pregnancy weight gain and length of gestation.

Previous studies have reported a modest U-shaped relation between total gestational weight gain and preterm delivery, where both low and high weight gain are associated with increased risk.[1] As demonstrated by Hutcheon et al,[3] using a single measure of total weight gain at delivery can lead to a biased estimate of the risk of preterm, where low weight gain is ostensibly associated with increased risk, as women who delivered earlier had less time to gain weight. Some investigators have attempted to avoid this issue by calculating an average rate of weight gain or an adequacy ratio relative to the Institute of Medicine recommendations.[4–8] These methods, however, rely on additional assumptions concerning the weight gain trajectory and may not completely eliminate this potential source of bias.[3] One major issue with using a single measure of total weight gain as the exposure is that, among the women who deliver at term, some of the weight is gained after 37 weeks, when they are no longer at risk for preterm delivery.

We propose an alternative means to address the correlation between weight gain and gestational age at delivery by reframing the binary outcome of preterm (<37 vs. ≥37 weeks of gestation) as time to delivery (i.e., gestational age at delivery), and incorporating this semicontinuous outcome of interest into a survival framework. Studies of preterm delivery rarely use time-to-event analysis, despite its methodological advantages.[9–11] The survival approach has the additional advantage of discriminating week-specific delivery risk across the continuum of gestational age. This could prove particularly useful in light of recent research suggesting that neonatal morbidities are differential even within the "term"

Submitted 27 March 2015; accepted 19 October 2015.
From the Epidemiology Branch, Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Bethesda, MD.
Supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, and by the Long-Range Research Initiative of the American Chemistry Council.
The authors report no conflicts of interest.

Correspondence: Enrique F. Schisterman, Epidemiology Branch, Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, 6100 Executive Blvd., Bethesda, MD 20852. E-mail: schistee@mail.nih.gov.

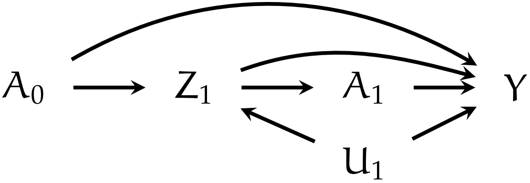
ISSN: 1044-3983/16/2702-0182
DOI: 10.1097/EDE.0000000000000413

182 | www.epidem.com *Epidemiology* • Volume 27, Number 2, March 2016

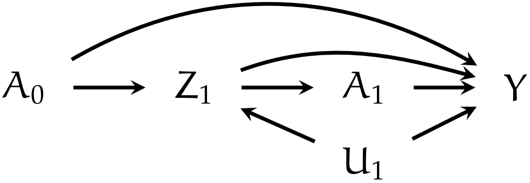
Analytic Complications Due to Time

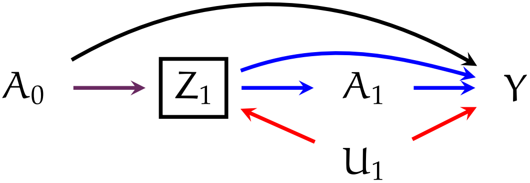
- Censoring & Competing Risks
- Correlated / Clustered Outcomes
- Left and Right Truncation
- Time-Dependent Confounding / Interaction
- Confounding by Time-Scale**

Time-Dependent Confounding (simplified)



Time-Dependent Confounding (simplified)





Time-Dependent Confounding: Examples

Overall effect of iron supplementation (A) during pregnancy on anemia at delivery (Y) confounded by hemoglobin and serum ferritin concentrations (Z; Bodnar 2004).

Time-Dependent Confounding: Examples

Overall effect of iron supplementation (A) during pregnancy on anemia at delivery (Y) confounded by hemoglobin and serum ferritin concentrations (Z; Bodnar 2004).

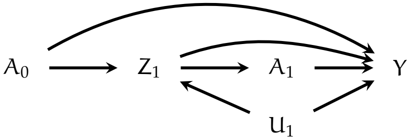
Overall effect of breastfeeding (A) on wheezing/atopy (Y) is confounded by infant weight gain (Z; Groenwold 2014).

Time-Dependent Confounding: Examples

Overall effect of iron supplementation (A) during pregnancy on anemia at delivery (Y) confounded by hemoglobin and serum ferritin concentrations (Z; Bodnar 2004).

Overall effect of breastfeeding (A) on wheezing/atopy (Y) is confounded by infant weight gain (Z; Groenwold 2014).

Overall effect of gestational weight gain (A) on infant mortality (Y) is confounded by gestational age at birth (Z; Mitchell 2015).



Time-Dependent Interaction: Examples

Does the effect of iron supplementation (A) in week j on anemia at delivery (Y) differ by past hemoglobin (Z) concentrations?

Time-Dependent Interaction: Examples

Does the effect of iron supplementation (A) in week j on anemia at delivery (Y) differ by past hemoglobin (Z) concentrations?

Does the effect of gestational weight gain (A) on perinatal mortality (Y) depend on the gestational week which it was gained (Z)?

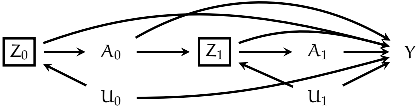
Time-Dependent Interaction: Examples

Does the effect of iron supplementation (A) in week j on anemia at delivery (Y) differ by past hemoglobin (Z) concentrations?

Does the effect of gestational weight gain (A) on perinatal mortality (Y) depend on the gestational week which it was gained (Z)?

Note how these differ from time-fixed (or baseline) interaction / effect modification.

Time-Dependent Interaction / Effect Modification

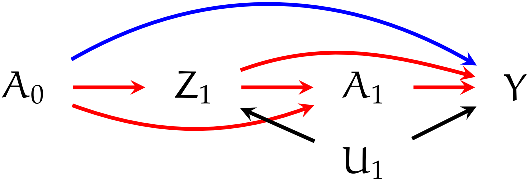


ID	t	A	Z	Y
1	0	0	0	119.65
1	1	1	0	119.65
2	0	0	0	87.29
2	1	0	1	87.29
3	0	1	1	137.72
3	1	1	1	137.72
4	0	0	1	105.28
4	1	0	1	105.28

We can't fit separate regression models for the effect of A on Y within levels of Z.

We can't include a main and interaction term between A and Z on Y.

Mediation



Direct

Indirect

How much of the effect of A_0 on Y is due to / independent of A_0's effect on A_1?

We can't quantify A_0's effect by simply adjusting for Z_1 and A_1.

The Meaning of Effect?

Thus far, we have used the word "effect" (overall, direct, indirect) informally.

This lack of formality can lead to vagueness, ambiguity, and problems with interpreting empirical results.

"Causal inference" seeks to address this.

Causal Inference & Potential Outcomes

- Causal Inference:

  A branch of scientific inquiry that combines identifiability assumptions with statistical methods to estimate causal (versus associational) effects

Causal Inference & Potential Outcomes

- Causal Inference:

  A branch of scientific inquiry that combines identifiability assumptions with statistical methods to estimate causal (versus associational) effects

- Potential Outcomes:

  The theoretical framework used to rigorously define what we mean by "causal effect"

Causal Inference & Potential Outcomes

- Causal Inference:

  A branch of scientific inquiry that combines identifiability assumptions with statistical methods to estimate causal (versus associational) effects

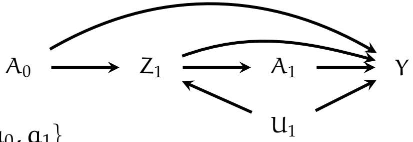
- Potential Outcomes:

  The theoretical framework used to rigorously define what we mean by "causal effect"

- Identifiability:

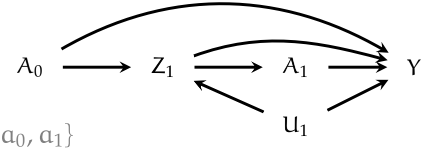
  An effect (defined via POs) is identifiable if it can be written as a function of the observed data

Potential Outcomes: ATE & ETT



- $Y^{\overline{a}}$: the outcome that would be observed if exposure were set to $\overline{a} = \{a_0, a_1\}$

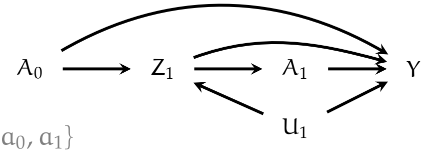
Potential Outcomes: ATE & ETT



- $Y^{\overline{a}}$: the outcome that would be observed if exposure were set to $\overline{a} = \{a_0, a_1\}$

- Different from the *observed* outcome.

Potential Outcomes: ATE & ETT



- $Y^{\overline{a}}$: the outcome that would be observed if exposure were set to $\overline{a} = \{a_0, a_1\}$

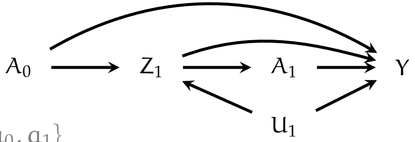
- Different from the *observed* outcome.

- Possible questions of interest:

$$E(Y^{1,1} - Y^{0,0}) \tag{ATE}$$

$$E(Y^{a_0,1} - Y^{a_0,0} \mid A_1 = 1) \tag{ETT}$$

Potential Outcomes: ATE & ETT



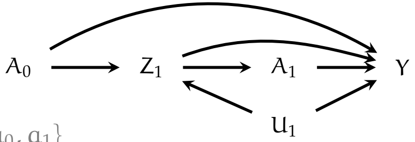
- $Y^{\overline{a}}$: the outcome that would be observed if exposure were set to $\overline{a} = \{a_0, a_1\}$
- Different from the *observed* outcome.
- Possible questions of interest:

$$E(Y^{1,1} - Y^{0,0}) \tag{ATE}$$

$$E(Y^{a_0,1} - Y^{a_0,0} \mid A_1 = 1) \tag{ETT}$$

- ATE: What is the average difference in POs if everyone received $\overline{a} = \{1, 1\}$ versus $\overline{a} = \{0, 0\}$?

Potential Outcomes: ATE & ETT



- $Y^{\overline{a}}$: the outcome that would be observed if exposure were set to $\overline{a} = \{a_0, a_1\}$
- Different from the *observed* outcome.
- Possible questions of interest:

$$E(Y^{1,1} - Y^{0,0}) \tag{ATE}$$

$$E(Y^{a_0,1} - Y^{a_0,0} \mid A_1 = 1) \tag{ETT}$$

- ATE: What is the average difference in POs if everyone received $\overline{a} = \{1, 1\}$ versus $\overline{a} = \{0, 0\}$?
- ETT: What is the average difference in POs if, among those who actually received A_1, everyone took A_1 versus no one took A_1?

Effect of Treatment on the Treated

- ETT/ATE is specific to a particular treatment/population.
- Under homogeneous treatment, ATE and ETT are the same.
- ATE averages over all units (including those very unlikely to be treated) & thus targets external validity.
- ETT measures the “biological impact” of a particular treatment
- Refer to handout on ATE v ETT for an example.

Potential Outcomes: The Fundamental Problem of Causal Inference

In general, it is impossible to observe different potential outcomes on the same individual and, therefore, impossible to observe the effect (ATE or ETT) of A on the outcome.

Takeaway: for a given individual, at least one potential outcome is always missing.

This is the FPCI.

Causal inference is about how we can (best) impute summaries of these missing potential outcomes.

G Methods

- Introduced by Robins ~ 1980s-1990s

G Methods

- Introduced by Robins $\sim$ 1980s-1990s
- Consist of three methods:

G Methods

- Introduced by Robins ~ 1980s-1990s
- Consist of three methods:
    - The (parametric) g formula

G Methods

- Introduced by Robins $\sim$ 1980s-1990s
- Consist of three methods:
  - The (parametric) g formula
  - Inverse probability weighted marginal structural models

G Methods

- Introduced by Robins ~ 1980s-1990s
- Consist of three methods:
  - The (parametric) g formula
  - Inverse probability weighted marginal structural models
  - G estimation of a structural nested model

G Methods

- Introduced by Robins $\sim$ 1980s-1990s
- Consist of three methods:
  - The (parametric) g formula
  - Inverse probability weighted marginal structural models
  - G estimation of a structural nested model
- The parametric g formula requires a model for *everything*

G Methods

- Introduced by Robins ~ 1980s-1990s
- Consist of three methods:
  - The (parametric) g formula
  - Inverse probability weighted marginal structural models
  - G estimation of a structural nested model
- The parametric g formula requires a model for *everything*
- IPW MSMs require a model for the exposure

G Methods

- Introduced by Robins ~ 1980s-1990s
- Consist of three methods:
  - The (parametric) g formula
  - Inverse probability weighted marginal structural models
  - G estimation of a structural nested model
- The parametric g formula requires a model for *everything*
- IPW MSMs require a model for the exposure
- We will focus today on g estimation and structural nested models

A Taxonomy of Structural Nested Models

There are different kinds of structural nested models:

- SN Mean Models
- SN Distribution Models

SNMM:

- Linear
- Log-linear
- Cumulative FT

SNDM:

- Linear
- Log-linear
- Accelerated FT

A Taxonomy of Structural Nested Models

There are different kinds of structural nested models:

- SN Mean Models
- SN Distribution Models

SNMM:

- Linear
- Log-linear
- Cumulative FT

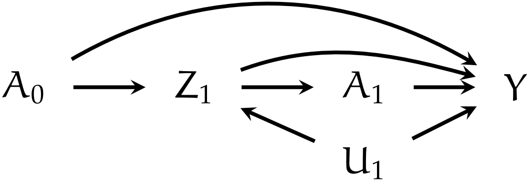
SNDM:

- Linear
- Log-linear
- Accelerated FT

When the mean does not adequately summarize the data, or interest lies in other components of the outcome distribution

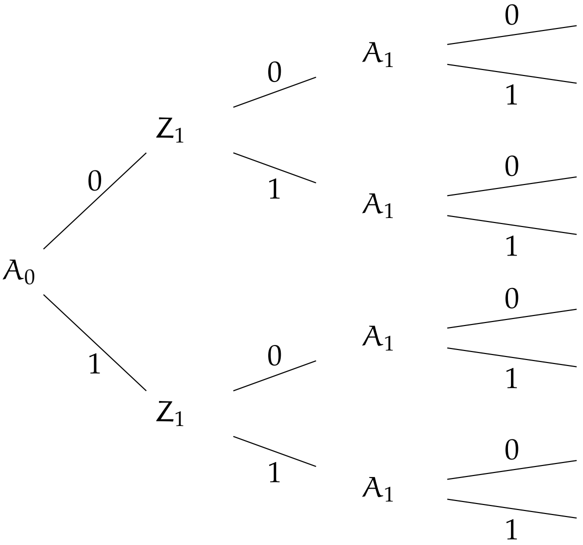
Working Example 1

- A_0, A_1: HAART at second and third trimester
- Z_1: HIV viral load at end of second trimester
- Y: CD4 count at end of third trimester



Any Questions?

Working Example I



Y	N
87.29	209, 271
112.11	93, 779
119.65	60, 657
144.84	136, 293
105.28	134, 781
130.18	60, 789
137.72	93, 903
162.83	210, 527

Working Example I

Row	A_0	Z_1	A_1	N	Y
1	0	1	0	60,657	119.65
2	0	1	1	136,293	144.84
3	0	0	0	209,271	87.29
4	0	0	1	93,779	112.11
5	1	1	0	93,903	137.72
6	1	1	1	210,527	162.83
7	1	0	0	134,781	105.28
8	1	0	1	60,789	130.18

Background Assumptions

- Non-Informative Censoring/Loss to Follow-up
- Missing Data Completely at Random (MCAR)
- No Measurement Error

Structural Nested Mean Model

$$E\left[Y^{a_0,0} - Y^{0,0} \mid A_0 = a_0\right] = \psi_0 a_0$$
$$E\left[Y^{a_0,a_1} - Y^{a_0,0} \mid A_0 = a_0, A_1 = a_1, Z_1\right] = \psi_1 a_1 + \psi_2 a_1 Z_1$$

- Structural: model for contrast of counterfactual outcomes
- Nested: counterfactual contrast nested in (conditional on) levels of A_0, and A_0, A_1, Z_1
- ψ quantifies the ETT: effect of treatment on the treated at time t, and then no treatment after that
- We can use g estimation to estimate ψ

Structural Nested Mean Model: Interpretation

- $\psi_1 + \psi_2$: The effect of HAART in 3^{rd} trimester ($A_1 = 1$) among those who actually received it and with high viral load at end of 2^{nd} trimester ($Z_1 = 1$).

- ψ_1: The effect of HAART in 3^{rd} trimester ($A_1 = 1$) among those who actually received it and with low viral load at end of 2^{nd} trimester ($Z_1 = 0$).

- ψ_0: The effect of HAART in 2^{nd} trimester ($A_0 = 1$) among those who actually received it, and no HAART in the 2^{nd} trimester ($A_1 = 0$).

These parameters (denoted "psi") quantify our causal effects of interest. Because of the FPCI, we can only estimate them under certain assumptions. These assumptions will be demonstrated in the example.

G Estimation of a SNMM: "By Hand"

The goal is to fill the last two columns of this table. We do this by assumption.

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65		
2	0	1	1	136,293	144.84		
3	0	0	0	209,271	87.29		
4	0	0	1	93,779	112.11		
5	1	1	0	93,903	137.72		
6	1	1	1	210,527	162.83		
7	1	0	0	134,781	105.28		
8	1	0	1	60,789	130.18		

G Estimation of a SNMM: "By Hand"

Counterfactual consistency:

- The potential outcome under the observed exposure is the observed outcome.

$$Y^{\{A_0,A_1\}} = Y,$$

where (capital) A_0, A_1 denotes the observed exposure at time zero and one.

G Estimation of a SNMM: “By Hand”

Step 1: Start filling in table by consistency assumption

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65		
2	0	1	1	136,293	144.84		
3	0	0	0	209,271	87.29		
4	0	0	1	93,779	112.11		
5	1	1	0	93,903	137.72		
6	1	1	1	210,527	162.83		
7	1	0	0	134,781	105.28		
8	1	0	1	60,789	130.18		

G Estimation of a SNMM: “By Hand”

Step 1: Start filling in table by consistency assumption

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65	119.65	119.65
2	0	1	1	136,293	144.84		
3	0	0	0	209,271	87.29		
4	0	0	1	93,779	112.11		
5	1	1	0	93,903	137.72		
6	1	1	1	210,527	162.83		
7	1	0	0	134,781	105.28		
8	1	0	1	60,789	130.18		

G Estimation of a SNMM: "By Hand"

Step 1: Start filling in table by consistency assumption

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65	119.65	119.65
2	0	1	1	136,293	144.84		
3	0	0	0	209,271	87.29	87.29	87.29
4	0	0	1	93,779	112.11		
5	1	1	0	93,903	137.72		
6	1	1	1	210,527	162.83		
7	1	0	0	134,781	105.28		
8	1	0	1	60,789	130.18		

G Estimation of a SNMM: “By Hand”

Step 1: Start filling in table by consistency assumption

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65	119.65	119.65
2	0	1	1	136,293	144.84		
3	0	0	0	209,271	87.29	87.29	87.29
4	0	0	1	93,779	112.11		
5	1	1	0	93,903	137.72	137.72	
6	1	1	1	210,527	162.83		
7	1	0	0	134,781	105.28		
8	1	0	1	60,789	130.18		

G Estimation of a SNMM: “By Hand”

Step 1: Start filling in table by consistency assumption

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65	119.65	119.65
2	0	1	1	136,293	144.84		
3	0	0	0	209,271	87.29	87.29	87.29
4	0	0	1	93,779	112.11		
5	1	1	0	93,903	137.72	137.72	
6	1	1	1	210,527	162.83		
7	1	0	0	134,781	105.28	105.28	
8	1	0	1	60,789	130.18		

G Estimation of a SNMM: “By Hand”

Step 1: Start filling in table by consistency assumption

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65	119.65	119.65
2	0	1	1	136,293	144.84		
3	0	0	0	209,271	87.29	87.29	87.29
4	0	0	1	93,779	112.11		
5	1	1	0	93,903	137.72	137.72	
6	1	1	1	210,527	162.83		
7	1	0	0	134,781	105.28	105.28	
8	1	0	1	60,789	130.18		

G Estimation of a SNMM: “By Hand”

Step 1: Start filling in table by consistency assumption

$$E\left[Y^{a_0,0} - Y^{0,0} \mid A_0 = a_0\right] = \psi_0 \; a_0$$

$$E\left[Y^{a_0,a_1} - Y^{a_0,0} \mid A_0 = a_0, A_1 = a_1, Z_1\right] = \psi_1 a_1 + \psi_2 a_1 Z_1$$

This is the effect of a_0. If we subtract it from $Y^{A_0,0}$, we get $Y^{0,0}$.

If this model is correct, we can use it to continue filling the table.

- Correct parametrically (model is unsaturated)
- Correct causally

G Estimation of a SNMM: “By Hand”

Step 1: Continue filling by consistency + correct model specification

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65	119.65	119.65
2	0	1	1	136,293	144.84		
3	0	0	0	209,271	87.29	87.29	87.29
4	0	0	1	93,779	112.11		
5	1	1	0	93,903	137.72	137.72	
6	1	1	1	210,527	162.83		
7	1	0	0	134,781	105.28	105.28	$105.28 - \psi_0$
8	1	0	1	60,789	130.18		

$$E\left[Y^{a_0,0} - Y^{0,0} \mid A_0 = a_0\right] = \psi_0 a_0$$

$$E\left[Y^{a_0,a_1} - Y^{a_0,0} \mid A_0 = a_0, A_1 = a_1, Z_1\right] = \psi_1 a_1 + \psi_2 a_1 Z_1$$

G Estimation of a SNMM: "By Hand"

Step 1: Continue filling by consistency + correct model specification

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65	119.65	119.65
2	0	1	1	136,293	144.84	$144.84 - \psi_1 - \psi_2$	$144.84 - \psi_1 - \psi_2$
3	0	0	0	209,271	87.29	87.29	87.29
4	0	0	1	93,779	112.11	$112.11 - \psi_1$	$112.11 - \psi_1$
5	1	1	0	93,903	137.72	137.72	$137.72 - \psi_0$
6	1	1	1	210,527	162.83	$162.83 - \psi_1 - \psi_2$	$162.83 - \psi_0 - \psi_1 - \psi_2$
7	1	0	0	134,781	105.28	105.28	$105.28 - \psi_0$
8	1	0	1	60,789	130.18	$130.18 - \psi_1$	$130.18 - \psi_0 - \psi_1$

$$E\left[Y^{a_0,0} - Y^{0,0} \mid A_0 = a_0\right] = \psi_0 a_0$$
$$E\left[Y^{a_0,a_1} - Y^{a_0,0} \mid A_0 = a_0, A_1 = a_1, Z_1\right] = \psi_1 a_1 + \psi_2 a_1 Z_1$$

G Estimation of a SNMM: "By Hand"

Exchangeability implies:

- $Y^{\{0,0\}} \coprod A_0$ (Marginal)
- $Y^{\{0,0\}} \coprod A_1 \mid A_0, Z_1$ (Conditional)

Therefore, for a given unique strata of $\{A_0, Z_1\}$, the mean of $Y^{0,0}$ among those with $A_1 = 0$ is equal to the mean of $Y^{0,0}$ among those with $A_1 = 1$

Exposure is independent of the potential outcomes

G Estimation of a SNMM: “By Hand”

Step 1: Solve for parameters by exchangeability

Row	A_0	Z_1	A_1	N	Y	$Y^{\{A_0,0\}}$	$Y^{\{0,0\}}$
1	0	1	0	60,657	119.65	119.65	119.65
2	0	1	1	136,293	144.84	$144.84 - \psi_1 - \psi_2$	$144.84 - \psi_1 - \psi_2$
3	0	0	0	209,271	87.29	87.29	87.29
4	0	0	1	93,779	112.11	$112.11 - \psi_1$	$112.11 - \psi_1$
5	1	1	0	93,903	137.72	137.72	$137.72 - \psi_0$
6	1	1	1	210,527	162.83	$162.83 - \psi_1 - \psi_2$	$162.83 - \psi_0 - \psi_1 - \psi_2$
7	1	0	0	134,781	105.28	105.28	$105.28 - \psi_0$
8	1	0	1	60,789	130.18	$130.18 - \psi_1$	$130.18 - \psi_0 - \psi_1$

G Estimation of a SNMM: “By Hand”

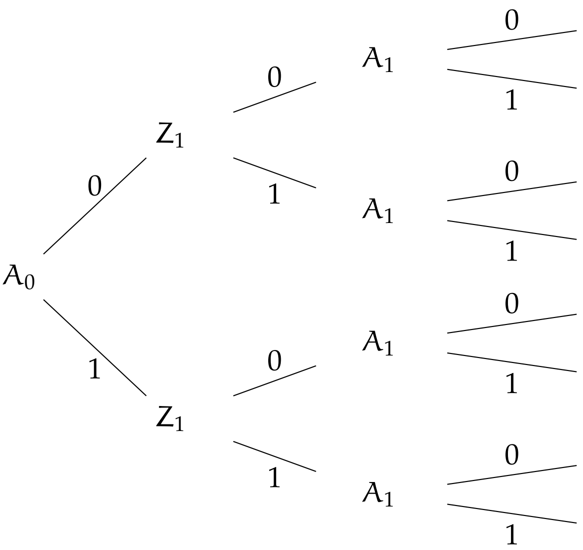
Results:

Parameter	Estimate
ψ_0	25.0
ψ_1	25.0
ψ_2	0

Among those who received HAART:

- Taking HAART in the second or third trimester increases CD4 by 25 cells/mm^3.
- The third trimester effect is constant across levels of HIV viral load (high/low).

Working Example I



Y	N
87.29	209, 271
112.11	93, 779
119.65	60, 657
144.84	136, 293
105.28	134, 781
130.18	60, 789
137.72	93, 903
162.83	210, 527

(modified) G Estimation of a SNMM

Let's assume we know $\psi_2 = 0$.

1. Estimate propensity score for A:

$$\pi_{A_1} = \{1 + \exp[-(\alpha_0 + \alpha_1 Z_1 + \alpha_2 A_0)]\}^{-1}$$
$$\pi_{A_0} = \{1 + \exp[-(\beta_0)]\}^{-1}$$

(modified) G Estimation of a SNMM

Let's assume we know $\psi_2 = 0$.

1. Estimate propensity score for A:

$$\pi_{A_1} = \{1 + \exp[-(\alpha_0 + \alpha_1 Z_1 + \alpha_2 A_0)]\}^{-1}$$
$$\pi_{A_0} = \{1 + \exp[-(\beta_0)]\}^{-1}$$

2. Estimate ψ_1 by fitting a linear regression model for Y, replacing A_1 with r_{A_1} and adding π_{A_1}:

$$E(Y \mid \hat{r}_{A_1}, A_0, Z_1, \hat{\pi}_{A_1}) = \psi_1 \hat{r}_{A_1} + \gamma_0 + \gamma_1 A_0 + \gamma_2 Z_1 + \delta_1 \hat{\pi}_{A_1}$$

(modified) G Estimation of a SNMM

3. Removing the effect of A_1 from Y

$$\widetilde{Y} = Y - \hat{\psi}_1 A_1.$$

(modified) G Estimation of a SNMM

3. Removing the effect of A_1 from Y

$$\widetilde{Y} = Y - \hat{\psi}_1 A_1.$$

4. Regress $\widetilde{Y}$ against r_{A_0} and add π_{A_0}:

$$E(\widetilde{Y} \mid \hat{r}_{A_0}, \hat{\pi}_{A_0}) = \psi_0 \hat{r}_{A_0} + \gamma_{00} + \delta_{10} \hat{\pi}_{A_0}$$

(modified) G Estimation of a SNMM

This approach gives two chances to adjust for confounding:

- By modeling the exposure to obtain a propensity score (π_A) and the exposure residuals (r_A)
- By modeling the outcome via the regression model $E(Y \mid A_0, A_1, Z_1)$

This is known as double-robustness

(modified) G Estimation: Regression Based

regression_based.sas

```
*G Estimation OF A SNMM (SAS);
*MODIFIED G-ESTIMATION OF A SNMM: REGRESSION BASED APPROACH;
*fit propensity score models;
proc logistic data=a desc;
     freq n;
     model a1 = z1 a0 ;
     output out=a pred=pi_a1;
proc logistic data=a desc;
     freq n;
     model a0 = ;
     output out=a pred=pi_a0;
run;quit;run;
```

(modified) G Estimation: Regression Based

```
data a; set a;res_a1 = a1-pi_a1;res_a0 = a0-pi_a0;run;
*model psi_1;
proc reg data=a;
     freq n;
     model y = res_a1 z1 a0 pi_a1;
     ods output ParameterEstimates=parm1
          (where=(Variable="res_a1") keep=variable estimate);
run;quit;run;
*housekeeping;
data parm1;set parm1; merg=1;
rename estimate=psi11;drop variable;
run;
data a;set a;merg=1;
run;
```

(modified) G Estimation: Regression Based

```
*subtract a1 effect from y;
data b;
     merge a parm1;
     by merg;
     y_tilde = y - psi11*a1;
run;
*regress transformed outcome against a0;
proc reg data=b;
     freq n;
     model y_tilde = res_a0 pi_a0;
     ods output ParameterEstimates=parm0
          (where=(Variable="res_a0") keep=variable estimate);
run;quit;run;
```

(modified) G Estimation: Regression Based

Results:

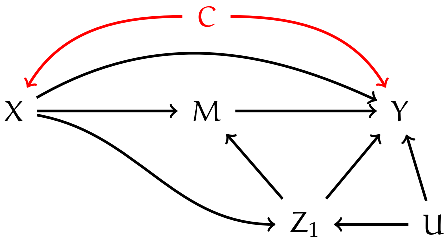
Parameter	Estimate
ψ_0	25.0
ψ_1	25.0

Among those who received HAART:

- Taking HAART in the second or third trimester increases CD4 by 25 cells/mm^3.

(modified) G Estimation: Mediation Analysis

- X: 2^{nd} trimester systolic BP.
- M: 3^{rd} trimester systolic BP.
- Z_1: Dietary+Exercise intervention.
- Y: Fetal or infant death.



Log-Linear SNMM.

Structural Nested Mean Model: Mediation Analysis

$$E\left[Y^{x,0} - Y^{0,0} \mid X = x\right] = \psi_0 x$$
$$E\left[Y^{x,m} - Y^{x,0} \mid X = x, M = m, Z_1\right] = \psi_1 m + \psi_2 m Z_1$$

- $\psi_1 + \psi_2$: The effect of a unit increase in 3^{rd} trimester BP with prior exercise/diet intervention.
- ψ_1: The effect of a unit increase in 3^{rd} trimester BP without prior exercise/diet intervention.
- ψ_0: The controlled direct effect of unit increase in 2^{nd} trimester BP (with 3^{nd} trimester BP fixed at zero value.

(modified) G Estimation: Mediation Analysis

- We will now fit log-linear SNMM for binary outcome
- Our goal is to estimate the CDE Risk Ratio
- Continuous exposure and mediator (linear regression to obtain propensity score)
- Same procedure, but must now use GLM with Gamma distribution and log link

(modified) G Estimation of a SNMM

1. Estimate propensity score for X and M and obtain residuals:

$$\begin{aligned}
\pi_X &= E[\alpha_0 + \alpha_1 C_{XY}] \\
\pi_M &= E[\beta_0 + \beta_1 C_{MY} + \beta_2 C_{XY} + \beta_3 X] \\
r_X &= X - \pi_X \\
r_M &= M - \pi_M
\end{aligned}$$

(modified) G Estimation of a SNMM

1. Estimate propensity score for X and M and obtain residuals:

$$\begin{aligned}
\pi_X &= \mathsf{E}[\alpha_0 + \alpha_1 C_{XY}] \\
\pi_M &= \mathsf{E}[\beta_0 + \beta_1 C_{MY} + \beta_2 C_{XY} + \beta_3 X] \\
r_X &= X - \pi_X \\
r_M &= M - \pi_M
\end{aligned}$$

2. Fit a log-linear Gamma GLM for Y, replacing M with r_M and adding π_M:

$$\begin{aligned}
\log \mathsf{E}(Y \mid X, \hat{r}_M, Z_1, \hat{\pi}_M) = \psi_1 \hat{r}_M + \psi_2 X \hat{r}_M \\
+ \gamma_{01} + \gamma_{11} Z_1 + \gamma_{21} X \\
+ \eta_{11} \hat{\pi}_M + \eta_{21} X \hat{\pi}_M
\end{aligned}$$

(modified) G Estimation of a SNMM

3. Removing the effect of M from Y

$$\widetilde{Y} = Y \times \exp(-\hat{\psi}_1 M - \hat{\psi}_2 XM).$$

(modified) G Estimation of a SNMM

3. Removing the effect of M from Y

$$\widetilde{Y} = Y \times \exp(-\hat{\psi}_1 M - \hat{\psi}_2 XM).$$

4. Regress $\widetilde{Y}$ against r_X, C, and add π_X:

$$\log \mathrm{E}(\widetilde{Y} \mid \hat{r}_X, C, \hat{\pi}_X) = \psi_0 \hat{r}_X + \beta_{00} + \beta_{10} C + \eta_{10} \hat{\pi}_X$$

(modified) G Estimation: Regression Based

snmmDR_LogLinear.sas

```
*DRG ESTIMATION OF A LOG LINEAR SNMM: REGRESSION BASED;
*propensity score;
proc reg data= a;
     model m = x c_xy c_my;
     output out=a pred=piM;
     ods select none;
run;quit;run;
proc reg data=a;
     model x = c_xy;
     output out=a pred=piX;
     ods select none;
run;quit;run;
```

(modified) G Estimation: Regression Based

- Have to multiply X by residual and PS for M
- What is `constant("small")`?

```
data a;
     set a;
     rM = m - piM;
     rX = x - piX;
     xrM = x*rM;
     xpiM = x*piM;
     y1 = y + constant("small");
run;
```

(modified) G Estimation: Regression Based

```
proc genmod data=a;
     class id;
     model y1 = rM xrM piM xpiM x c_xy c_my
          / dist=gamma link=log;
     repeated subject=id / type=ind;
     ods output GEEEmpPEst=parm1
          (where=(parm="rM"|parm="xrM")keep = parm estimate);
run;quit;run;
*housekeeping;
proc transpose data=parm1 out=parm1(drop=_name_)
          prefix=psi_;id parm;run;
data parm1;set parm1;merg=1;
data a;set a;merg=1;run;
```

(modified) G Estimation: Regression Based

```
data b;
     merge a parm1;
     by merg;
     y1_tilde = y1*exp( - psi_rM*m - psi_xrM*xm);
proc genmod data=b;
     class id;
     model y1_tilde = rX piX c_xy / link=log dist=gamma;
     repeated subject=id / type=ind;
     ods output GEEEmpPEst=cde(where=(parm="rX"));
run;quit;run;
```

Results

$\mathsf{CDE} = 1.27$, 95% CI: 1.18, 1.37

- The risk of mortality due to the direct effect a one-unit systolic BP increase in the second trimester is 1.27 times the risk of no increase.
- Assumes linear dose-response, which can be relaxed using, e.g., polynomials

Concluding Remarks

- SNMs are useful for complex time-dependent confounding and scenarios, and questions related to time-dependent interaction.
- The regression-based approach presented here greatly facilitates implementation.
- Ideally, several modeling strategies targeting the same causal quantity of interest should be used in a given project.
- Plenty of user-friendly options are becoming increasingly available.

Structural Nested Models in Reproductive Epidemiology

Ashley I Naimi, PhD
ashley.naimi@pitt.edu

June 20 2016