Structural nested models and g-computation: policy and causal inference

$$
\begin{aligned}
\text{Risk}(\text{set } A = a) &= 1 - S(y_m^a) \\
&= Pr(T^a < k, max(\overline{Y}^a) = 1) \\
&= \sum_{k=1}^{m} \sum_{\overline{a}} \sum_{\overline{a}^*} \sum_{\overline{\ell}} \left[\begin{array}{l} Pr(Y_k = 1 | \overline{A}_k = \overline{a}_k, \overline{L}_k = \overline{\ell}_k, \overline{Y}_{k-1} = \overline{N}_k = \overline{C}_k = \overline{0}) \\ \prod_{j=1}^{k} \left[\begin{array}{l} Pr_d(A_j = a_j | \overline{A}_j^* = \overline{a}_j^*, \overline{A}_{j-1} = \overline{a}_{j-1}, \overline{L}_j = \overline{\ell}_j, \overline{Y}_{j-1} = \overline{N}_{j-1} = \overline{C}_{j-1} = \overline{0}) \\ Pr(A_j^* = a_j^* | \overline{A}_{j-1} = \overline{a}_{j-1}, \overline{L}_j = \overline{\ell}_j, \overline{Y}_{j-1} = \overline{N}_{j-1} = \overline{C}_{j-1} = \overline{0}) \\ Pr(L_j = \ell_j | \overline{A}_{j-1} = \overline{a}_{j-1}, \overline{L}_{j-1} = \overline{\ell}_{j-1}, \overline{Y}_{j-1} = \overline{N}_{j-1} = \overline{C}_{j-1} = \overline{0}) \\ Pr(Y_{j-1} = N_{j-1} = 0 | \overline{A}_{j-1} = \overline{a}_{j-1}, \overline{L}_{j-1} = \overline{\ell}_{j-1}, \overline{Y}_{j-2} = \overline{N}_{j-2} = \overline{C}_{j-2} = \overline{0}) \end{array} \right] \end{array} \right]
\end{aligned}
$$

Alexander Keil, akeil@unc.edu

Dept. of Epidemiology
University of North Carolina at Chapel Hill

Materials

Course materials here:
`https://github.com/alexpkeil1/2018_ISEE_causal`

Or download directly:
`https://github.com/alexpkeil1/2018_ISEE_causal/archive/master.zip`

Objectives

1. Motivate use of the g-formula, structural nested models
2. Basic introduction to each approach
3. Follow multiple (simplified) applied examples
4. Get a resource list for learning more

Causal effect estimation in occupational and environmental epidemiology

[0] S. Greenland (2017) *Eur J Epidemiol*

Why causal?

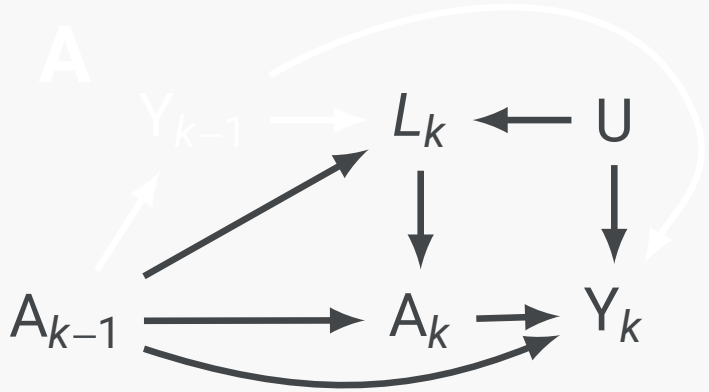
Why learn causal effect estimation?

- Bias (e.g. health worker survivor bias)
- Directly estimate public health actions

Why causal?

Healthy worker survivor bias

Healthy person = long employment = high exposure



time k

exposure A_k

employment status L_k

mortality Y_k

Why causal?

Healthy worker survivor bias

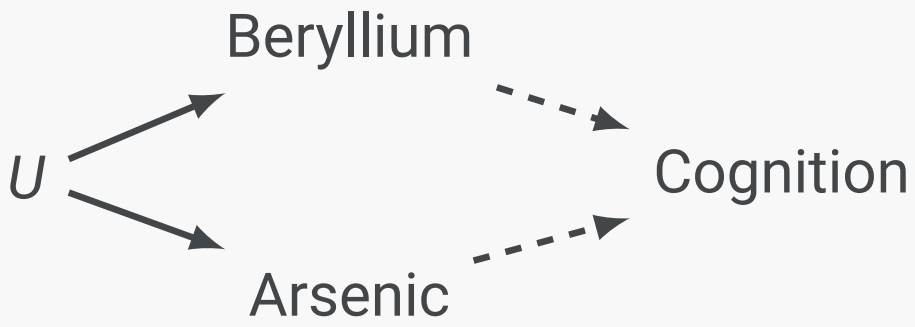
- In HWSB, L_k is an example of a time-varying confounder that is affected by prior exposure
- MSMs, g-computation, and structural nested models can address such bias

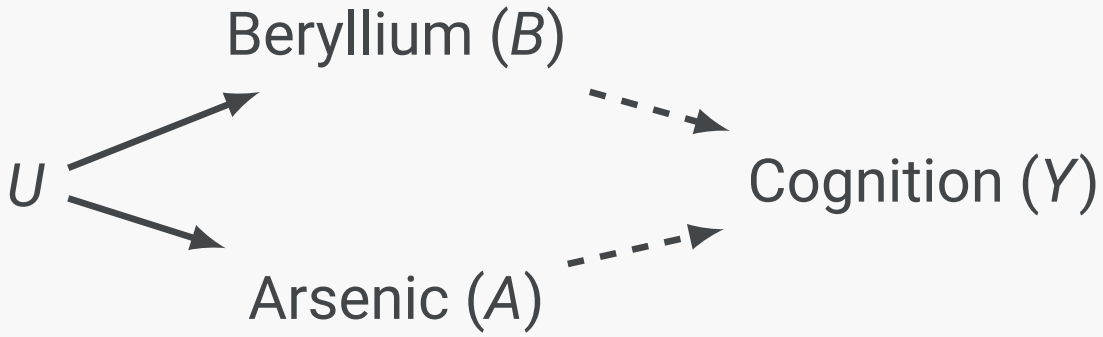
Why causal?

Directly estimate public health actions

Say we have two correlated ambient (airborne) exposures and one outcome and we would like to quantify their relationship

This is how we think they are related





Most **approaches** we could use to learn about arsenic, beryllium, and cognition involve a linear model such as:

$$Y_i = \beta_0 + \beta_1 A_i + \beta_2 B_i + \epsilon_i$$

Most **applications** involve directly interpreting β_1, β_2 (e.g. expected loss in IQ per $10 ng/m^3$ increase in Beryllium, holding Arsenic fixed)

A model such as

$$Y_i = \beta_0 + \beta_1 A_i + \beta_2 B_i + \epsilon_i$$

is causal[1] in that we have controlled confounding by co-exposure, but interpreting β_1, β_2 directly leads to our inference being driven by the model, rather than the study questions of inference[2]

[1] 'causal' doesn't always mean 'useful'

[2] and β_1, β_2 bear little resemblance to experimentally verifiable effects - it is rare that we could directly intervene on one air pollutant without affecting others

Joint effects, public health

Simply interpreting β_1, β_2 may not reflect information among exposures as well as joint effects, but which joint effect?

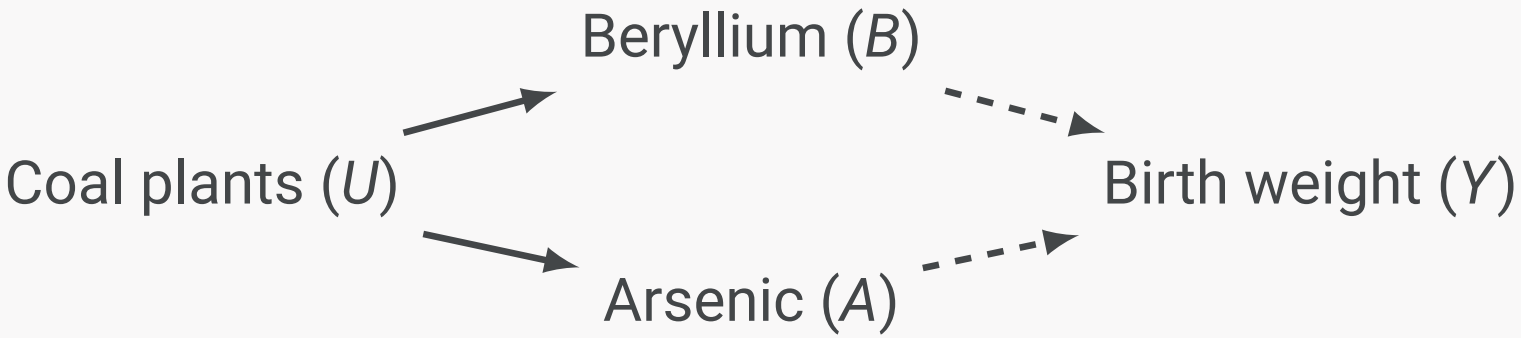
We could consider why exposures are correlated.

Joint effects, public health

Simply interpreting β_1, β_2 may not reflect information among exposures as well as joint effects, but which joint effect?

We could consider why exposures are correlated.

We could envision a public health intervention on the source(s) of co-exposures.

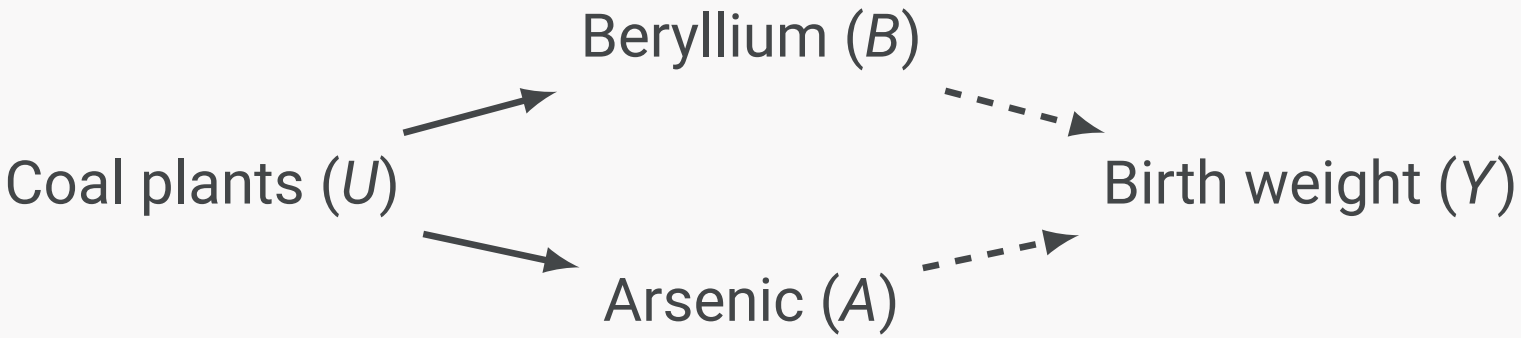


Joint effects, public health

Simply interpreting β_1, β_2 may not reflect information among exposures as well as joint effects, but which joint effect?

We could consider why exposures are correlated.

We could envision a public health intervention on the source(s) of co-exposures.



Causal inference methods can help here, too

Why causal?

Learning examples[3]

- Bias (e.g. health worker survivor bias)
- Directly estimate public health actions

[3] Using simulated data to emulate published or ongoing examples. Code and data for these analyses are available as part of the workshop materials.

Why causal?

Learning examples[3]

- Bias (e.g. health worker survivor bias)
- Directly estimate public health actions
    - 1. Exploring effects of coal plant emissions on cognitive functioning using the **parametric g-formula**

[3] Using simulated data to emulate published or ongoing examples. Code and data for these analyses are available as part of the workshop materials.

Why causal?

Learning examples[4]

- Bias (e.g. health worker survivor bias)
- Directly estimate public health actions
  - 1. Exploring effects of coal plant emissions on cognitive functioning using the **parametric g-formula**
  - 2. Estimating effects on mortality of hypothetical new occupational standards for radon exposure in an occupational cohort using the **parametric g-formula**[3]

[3] There is also an example of a similar analysis using and **structural nested models** in the workshop code

[4] Using simulated data to emulate published or ongoing examples. Code and data for these analyses are available as part of the workshop materials.

Why causal?

Learning examples[4]

- Bias (e.g. health worker survivor bias)
  - 3. Estimating exposure-response functions of cumulative radon exposures on mortality controlling for healthy worker survivor bias using **structural nested models**
- Directly estimate public health actions
  - 1. Exploring effects of coal plant emissions on cognitive functioning using the **parametric g-formula**
  - 2. Estimating effects on mortality of hypothetical new occupational standards for radon exposure in an occupational cohort using the **parametric g-formula**[3]

---

[3] There is also an example of a similar analysis using and **structural nested models** in the workshop code

[4] Using simulated data to emulate published or ongoing examples. Code and data for these analyses are available as part of the workshop materials.

Note

This talk is about why and how to use the g-formula and/or structural nested models, but not when they should not be used

We presume all causal assumptions are met (exchangeability, consistency, positivity when necessary) and only briefly discuss model specification.

The g-formula: vaguely

G-computation algorithm formula

Using data and assumptions to make inference about effects of interventions or treatments

For example: causal risk difference comparing always versus never exposed

The g-formula: utility

Useful for

- Target parameters that don’t come from a model
- Population level impacts
- Complex, longitudinal data
- Dynamic exposure, treatment regimes
- Heuristic tool: computer coding directly tied to potential outcomes

Notation (examples)

Random variables: e.g. B, A, L

Realizations: b, a, ℓ

Values at time k: B_k, a_k, ℓ_k

History through time k: $\overline{B}_k, \overline{a}_k, \overline{\ell}_k$

Parameters: β, θ, γ

Probability/dist: $Pr(B = 1), p(x), p(y^g)$

Exposure: A, A_k, a, a_k

Confounder(s): L, L_k, ℓ, ℓ_k

Outcome: Y, Y_k, y, y_k

Potential outcomes: $Y^1, Y^g, Y^1_k, Y^{\bar{g}}_k$

Definition

Regime: "Set" exposure A_k to some value g. Note that g can be static or may depend on k, as well as prior covariates: $g = g_k(\bar{L}_k)$ (dynamic regime)

We will skip ahead if Dr. Kaufman had time to give a basic introduction to g-computation, but the following introductory slides are for your reference.

The g-formula

Basic building block

Law of total probability

$$Pr(A) = \sum_{b} Pr(A|B = b)Pr(B = b)$$

The g-formula

Basic building block

Law of total probability

$$Pr(A) = \sum_b Pr(A|B = b)Pr(B = b)$$

Direct standardization

$$Pr(Y = 1) = \sum_a Pr(Y = 1|Age = a)Pr(Age = a)$$

The g-formula

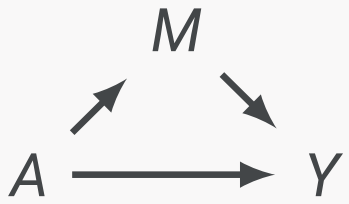
Basic building block

Law of total probability, part 2

$$Pr(A|C = c) = \sum_b Pr(A|B = b, C = c)Pr(B = b|C = c)$$

Regression v. standardization

Bias via adjustment



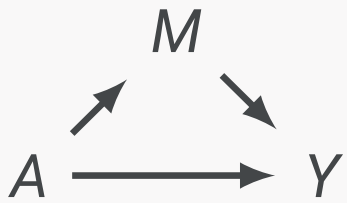
Given a logistic regression model:

$$Pr(Y = 1|M, A; \beta) = expit(\beta_0 + \beta_1 m + \beta_2 a)$$

Does the $E(Y|A = 1) = expit(\beta_0 + \beta_2)$ (or $expit(\beta_0 + \beta_1 + \beta_2)$)?

Regression v. standardization

No bias via standardization

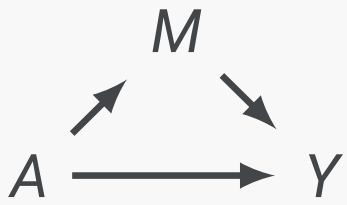


Using standardization:
Does $E(Y|X = 1) =$

$$\sum_{m \in \mathcal{M}} Pr(Y = 1|M = m, A = 1)Pr(M = m|A = 1) \ ?$$

Regression v. standardization

Why does this matter?

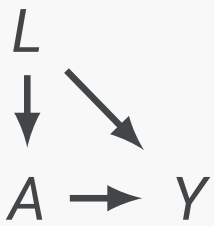


M could be: a future value of exposure, a confounder of a future value of exposure or some other quantity that is important to inference. We don't want regression adjustment/stratification.

More simply: we can use the g-formula to combine non-causal quantities to get a causal contrast

- Standardization: allows consideration of causal intermediates as covariates
- Law of total probability = basic building block for causal inference when covariates and exposures influence each other

The g-formula: time-fixed



Evaluate at $A = g$

$$Pr\left(Y^g\right) = \sum_{\ell} Pr\left(Y^g \middle| L = \ell\right) Pr\left(L = \ell\right) \qquad \text{Law of total probability}$$

$$= \sum_{\ell} Pr\left(Y^g \middle| A = g, L = \ell\right) Pr\left(L = \ell\right) \qquad \text{Exchangeability}$$

$$= \sum_{\ell} Pr\left(Y \middle| A = g, L = \ell\right) Pr\left(L = \ell\right) \qquad \text{Consistency, positivity}^5$$

[5] $f(g|l) > 0$ so that $Pr\left(Y \middle| A = g, L = \ell\right)$ exists

Example 1: time-fixed

L=0

		A		
		0	1	
Y	0	45	22	67
	1	30	28	58
		75	50	125

L=1

		A		
		0	1	
Y	0	28	10	38
	1	22	15	37
		50	25	75

Target parameter: Average causal risk difference

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

$$= Pr(Y^1 = 1) - Pr(Y^0 = 1)$$

L=0

		A = 0	A = 1	
Y	0	45	22	67
	1	30	28	58
		75	50	125

L=1

		A = 0	A = 1	
Y	0	28	10	38
	1	22	15	37
		50	25	75

Recall $Pr(Y^g) = \sum_\ell Pr\left(Y \middle| A = g, L = \ell\right) Pr\left(L = \ell\right)$

$Pr(L = 0)$ =125/200 = 0.625, $Pr(L = 1)$ =0.375

$Pr\left(Y = 1 \middle| A = 1, L = 0\right)$ =28/50 = 0.56

$Pr\left(Y = 1 \middle| A = 1, L = 1\right)$ =15/25 = 0.6

$Pr\left(Y = 1 \middle| A = 0, L = 0\right)$ =_____

$Pr\left(Y = 1 \middle| A = 0, L = 1\right)$ =_____

Click here for answers here

L=0

		A = 0	A = 1	
Y	0	45	22	67
	1	30	28	58
		75	50	125

L=1

		A = 0	A = 1	
Y	0	28	10	38
	1	22	15	37
		50	25	75

$$E(Y^1) = Pr\left(Y = 1 \middle| A = 1, L = 0\right) Pr(L = 0) + Pr\left(Y = 1 \middle| A = 1, L = 1\right) Pr(L = 1)$$
$$= 0.56 \times 0.625 + 0.6 \times 0.375 = 0.575$$

$$E(Y^0) = Pr\left(Y = 1 \middle| A = 0, L = 0\right) Pr(L = 0) + Pr\left(Y = 1 \middle| A = 0, L = 1\right) Pr(L = 1)$$
$$= \_\_\_\_ \times 0.625 + \_\_\_\_ \times 0.375$$
$$E(Y^1) - E(Y^0) = 0.575 - \_\_\_\_ = \_\_\_\_?$$

Click here for answers here

Basic algorithm - 1 time point[8]

G-formula for a static regime $set(A = g)$

- Start with distribution of observed data:
  $p(y, a, l) = p(y|a, \ell)p(a|\ell)p(\ell)$
- Replace $p(a|\ell)$ with a *degenerate* distribution $p_d(a|\ell)$ that is equal to 1 at $A = g$ and 0 everywhere else
- Marginalize over $p(\ell)$:
  $\int p(y|a, \ell)p_d(g|\ell)p(\ell)d\ell = \int p(y|g, \ell)p(\ell)d\ell$
- Note that $p(y|a, \ell)$ can be estimated via a regression model[6] $p(y|a, \ell, \beta)$, and marginalizing over $p(\ell)$ can be done by sample average of predictions from that model! EASY![7]

---

[6] When the regression model is not saturated, we call this the parametric g-formula

[7] J. M. Snowden S. Rose K. M. Mortimer (2011) *Am J Epidemiol*

[8] A is referred to as a "point exposure" in this case

Software note: I try to do things in SAS according to ’the SAS way’ and things in R according to ’the R way,’ so the code won’t match 1:1, but it helps ensure that the code is readable in both languages, and moderately efficient. I will demonstrate both when possible, but will otherwise show either R or SAS according to which more clearly addresses a point.

Example 1: coal plants and IQ

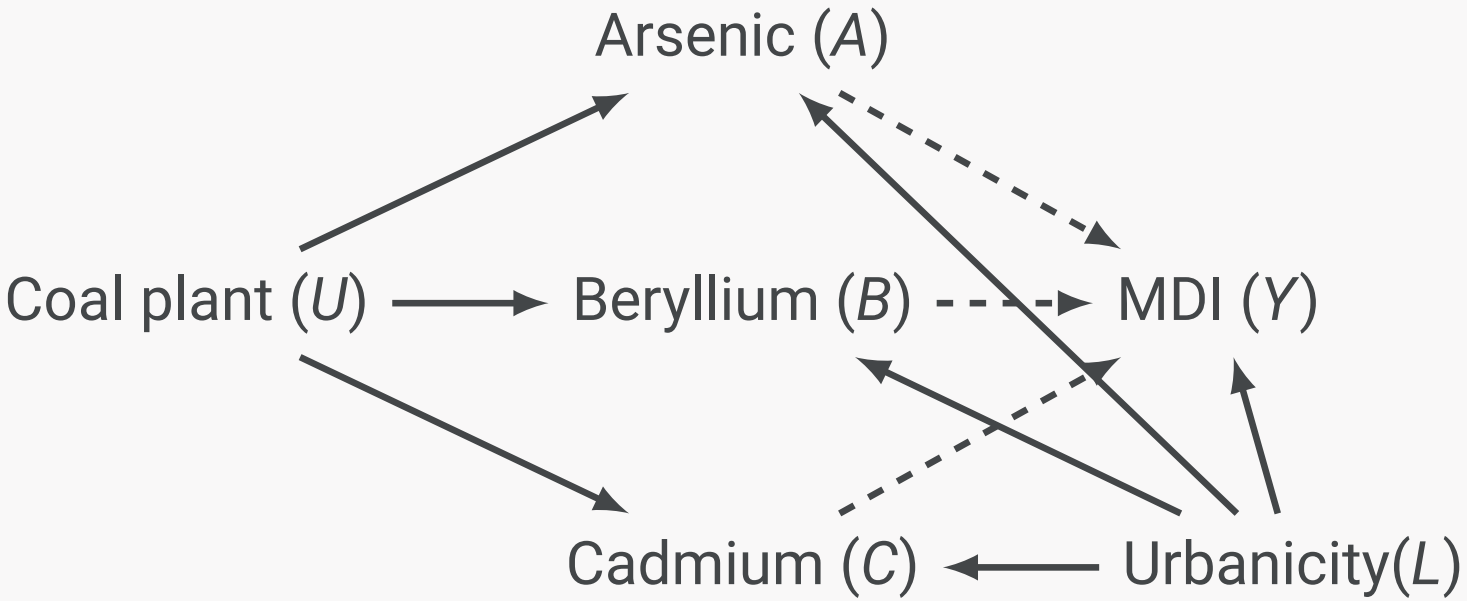
The parametric g-formula with point exposures

- Birth cohort[9] of 3,961 followed to 2 years of age in a US city
- Outcome: Mental Development Index measured at age 2
- Exposures: 3 metals known to be emitted from coal-fired power plants (As, Be, Cd) [annual mean ambient levels from birth to age 1, measured via passive monitoring]
- Confounder/modifier: Urbanicity

We have prior evidence that these 3 metals affect cognitive functioning in children, so we'd like to quantify their relationships with MDI in our cohort

[9] simulated

Co-pollutant confounding



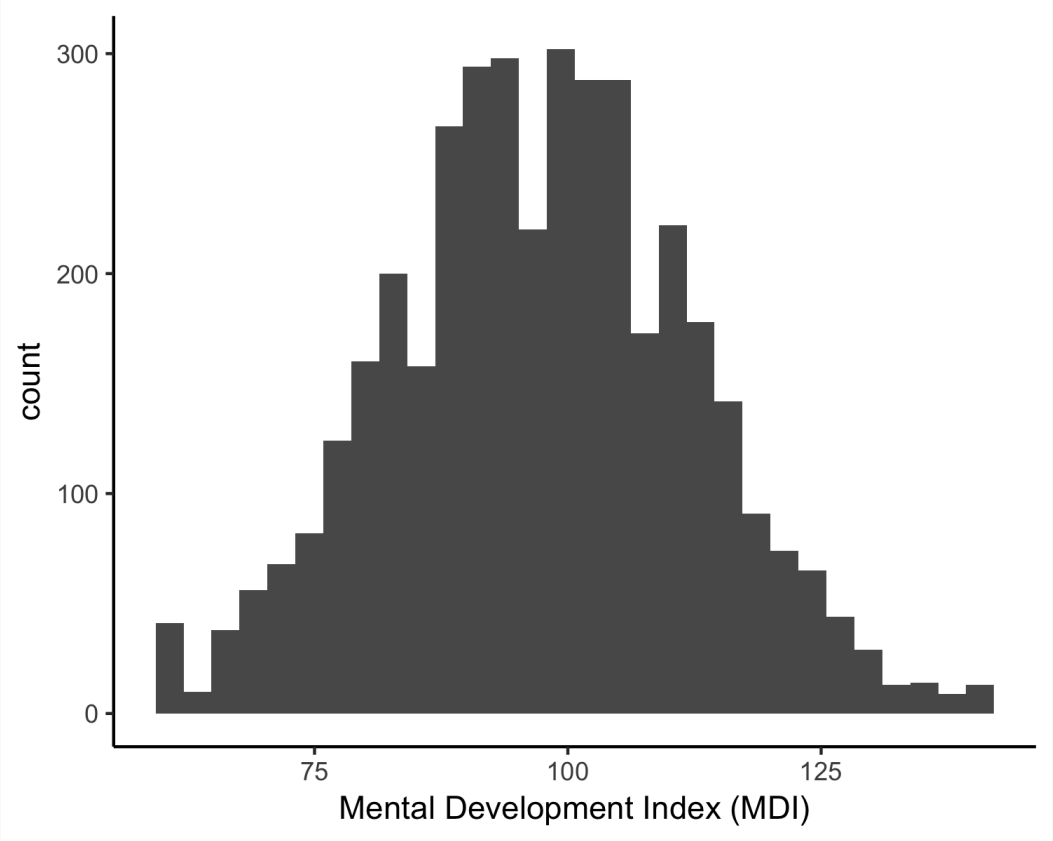
coalplant.csv

First 10 observations

id	as	be	cd	mdi	urbanicity
1	0.797	0.945	0.964	107	1
2	0.687	0.533	0.353	73	1
3	1.067	2.401	3.104	140	1
4	1.042	2.060	3.382	80	0
5	0.964	1.081	1.081	83	1
6	1.301	0.617	3.428	104	1
7	1.149	1.756	9.843	79	1
8	0.430	0.304	0.135	79	1
9	0.690	0.771	0.457	71	1
10	0.561	4.268	0.319	91	1

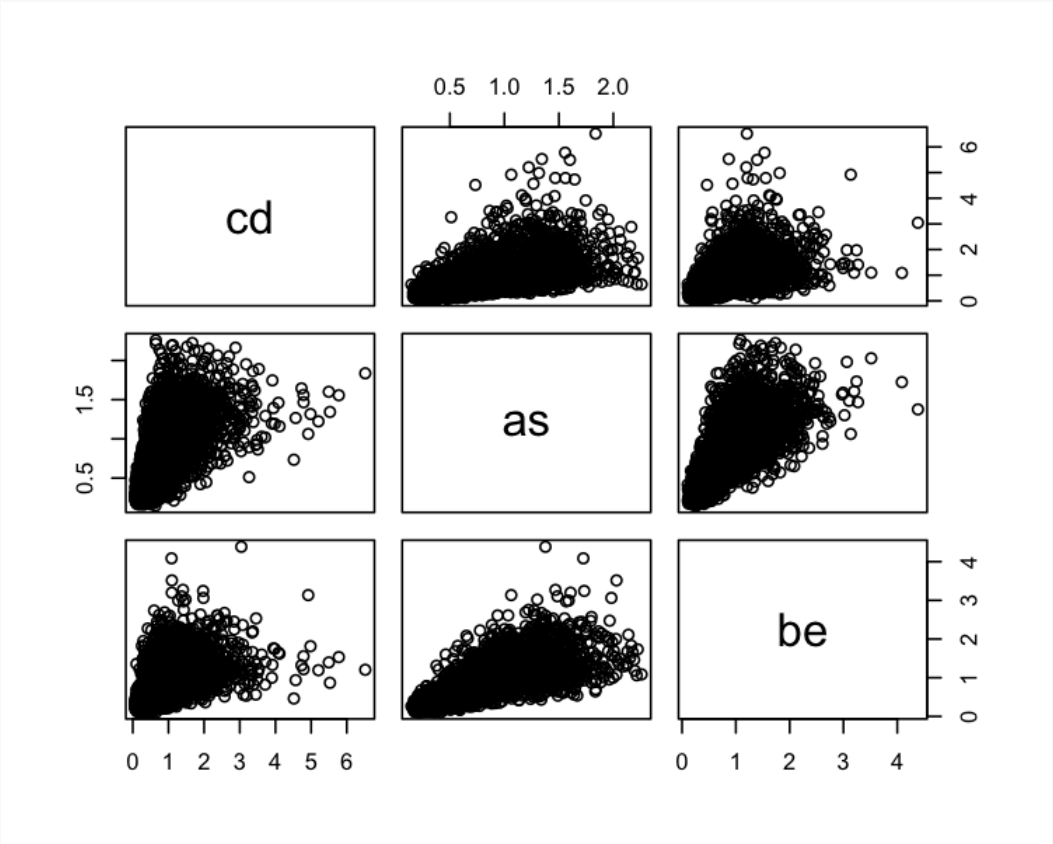
[9] Exposures measured in units of 10 ng/m^3 in air, MDI is raw score

MDI



Exposures

Correlation through a common source



Any realistic intervention on any of these metals would likely affect the other metals. Joint effect may therefore be of greatest interest.

Choice of joint effect

Percent emissions from local coal-fired power plant

From prior emissions data and consulting experts, we know that 91% of As, 96% of Be, and 45% of Cd ambient levels come from a local coal-fired power plant. One joint effect of interest would be reducing each of the exposures by the proportion expected if we had decommissioned the coal-fired power plant.

E.g. if a 1 year old girl had been exposed to 1 unit of Cd, under the intervention we would expect her to be exposed to only 0.55 units (1*(1-0.45))

$Y^{a,b,c}$ (“natural course”) vs. $Y^{a\times 0.09, b\times 0.04, c\times 0.55}$ (“shutdown”)

Other effects

Population attributable MDI differences

We also estimate population attributable MDI differences for each constituent pollutant[10], given as

$E(Y^{0,b,c}) - E(Y^{a,b,c})$ (As)
$E(Y^{a,0,c}) - E(Y^{a,b,c})$ (Be)
and
$E(Y^{a,b,0}) - E(Y^{a,b,c})$ (Cd)

[10] The g-formula gives straightforward way to estimate attributable fractions that are not subject to pitfalls of many other methods.

Step 1: modeling

Parametric g-formula for point-exposures

Need only $p(y|a, b, c, \ell)$, it's good to be flexible.[11] We chose a linear model with all first order product terms.

SAS:

```
PROC GENMOD DATA = coalplant;
  TITLE "Parametric g-formula model for point exposures";
  MODEL mdi = as be cd urbanicity as*urbanicity  be*urbanicity  cd*urbanicity  as*be  as*cd  be*cd;
  STORE obsmodel; *the STORE statement allows us to make predictions from this model in a later step;
```

R: (product terms expand automatically to include main terms)

```
mdimod = glm(mdi ~ as*urbanicity + be*urbanicity + cd*urbanicity + as*be + as*cd + be*cd, data=coalplant)
```

[11] Within reason. We will lose some efficiency, but gain

Step 1: modeling

Parametric g-formula for point-exposures

Need only $p(y|a, b, c, \ell)$, it's good to be flexible

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	100.9026	1.6409	97.6864	104.1188	3781.08	<.0001
as	1	-3.5976	3.2870	-10.0401	2.8448	1.20	0.2737
be	1	-3.2484	2.8949	-8.9224	2.4256	1.26	0.2618
cd	1	1.6525	2.1585	-2.5782	5.8831	0.59	0.4439
urbanicity	1	-0.0743	1.5535	-3.1192	2.9706	0.00	0.9618
as*urbanicity	1	3.3878	3.1549	-2.7957	9.5714	1.15	0.2829
be*urbanicity	1	-0.1697	2.5480	-5.1636	4.8242	0.00	0.9469
cd*urbanicity	1	-3.4989	1.7963	-7.0196	0.0218	3.79	0.0514
as*be	1	0.9772	1.4268	-1.8192	3.7737	0.47	0.4934
as*cd	1	-0.0953	1.1766	-2.4014	2.2108	0.01	0.9355
be*cd	1	0.0401	0.9097	-1.7429	1.8231	0.00	0.9649
Scale	1	14.7915	0.1662	14.4694	15.1208		

Step 1: modeling

Parametric g-formula for point-exposures

Often referred to as “Suggestive evidence” (AKA p-values>0.05)

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	100.9026	1.6409	97.6864	104.1188	3781.08	<.0001
as	1	-3.5976	3.2870	-10.0401	2.8448	1.20	0.2737
be	1	-3.2484	2.8949	-8.9224	2.4256	1.26	0.2618
cd	1	1.6525	2.1585	-2.5782	5.8831	0.59	0.4439
urbanicity	1	-0.0743	1.5535	-3.1192	2.9706	0.00	0.9618
as*urbanicity	1	3.3878	3.1549	-2.7957	9.5714	1.15	0.2829
be*urbanicity	1	-0.1697	2.5480	-5.1636	4.8242	0.00	0.9469
cd*urbanicity	1	-3.4989	1.7963	-7.0196	0.0218	3.79	0.0514
as*be	1	0.9772	1.4268	-1.8192	3.7737	0.47	0.4934
as*cd	1	-0.0953	1.1766	-2.4014	2.2108	0.01	0.9355
be*cd	1	0.0401	0.9097	-1.7429	1.8231	0.00	0.9649
Scale	1	14.7915	0.1662	14.4694	15.1208		

Step 1: modeling

Parametric g-formula for point-exposures

Most epidemiologic analyses don’t get past this point.
Coefficients are interpreted and then discussed.

Step 2: predictions

Parametric g-formula for point-exposures

We predict the expected outcome under each intervention, i.e. $E(Y^{a,b,c})$ for specific values of a, b, c

SAS:

```
DATA ints;
  SET coalplant;
  ARRAY a[3] cd as be;
  ARRAY _a[3] _cd _as _be;
  DO i = 1 TO 3;
    *store original values - they will be modified below, but we need to keep them;
    _a[i] = a[i];
  END;
  * no intervention = natural course;
  int = 'NC      ';
  OUTPUT;
  * decommission the coal fired power plant;
  int = 'No coal';
  cd=_cd*(1-0.5);as=_as*(1-0.96);be=_be*(1-0.91);
  OUTPUT;
  DO i = 1 TO 3;
    *restore original values;
    a[i] = _a[i];
  END;
  *eliminate cadmium without intervening on other pollutants;
  int = 'No Cd';
  cd=0;
  OUTPUT;
  cd = _cd;
  *eliminate cadmium without intervening on other pollutants;
  int = 'No As';
  as=0;
  OUTPUT;
  as = _as;
  *eliminate cadmium without intervening on other pollutants;
  int = 'No Be';
  be=0;
  OUTPUT;
* now make predictions based on these new exposures;
PROC PLM RESTORE=obsmodel;
  SCORE DATA = ints OUT=ints PRED = p_mdi;
RUN;

PROC MEANS DATA = ints MEAN;
  TITLE 'G-formula point estimates for mean MDI expected under each intervention';
  CLASS int;
  VAR p_mdi;
RUN;
```

R:

```
gformula_means <- function(ymodel, data, ...){
   # function to calculate mean MDI under a given intervention
   require('dplyr')
   # if only 'ymodel' and 'data' are supplied, then natural course is fit
  postinterventionX <- mutate(data,...)
  mean(predict(ymodel, newdata=postinterventionX))
}

#point estimates for mean MDI under each intervention
nc = gformula_means(mdimod, data=coalplant)
no_coal = gformula_means(mdimod, data=coalplant,
                         cd=cd*(1-0.5), as=as*(1-0.96), be=be*(1-0.91))
no_cd = gformula_means(mdimod, data=coalplant, cd=0)
no_as = gformula_means(mdimod, data=coalplant, as=0)
no_be = gformula_means(mdimod, data=coalplant, be=0)
# expected population mean MDI
print(c(nc=nc, no_coal=no_coal, no_cd=no_cd, no_as=no_as, no_be=no_be), 4)
```

Step 2: predictions

Parametric g-formula for point-exposures

Estimates of the predicted population mean MDI under each intervention:

Analysis Variable : p_mdi Predicted Value		
int	N Obs	Mean
NC	3961	97.3158293
No As	3961	97.0822896
No Be	3961	99.1578454
No Cd	3961	98.5691281
No coal	3961	99.9842099

Under causal assumptions, these are equal to counterfactual means under the hypothetical interventions.

Step 3: effect measures

Parametric g-formula for point-exposures

Code omitted from slides: to calculate the effect of decommissioning a plant, we subtract the mean expected MDI under the natural course (reference - the simulated act of no intervention) from the mean expected MDI under the intervention.

e.g. mean MDI difference = $E(Y^{a\times 0.09, b\times 0.04, c\times 0.55})$-$E(Y^{a,b,c})$ for estimating the effect of coal-plant decommissioning

Step 4: bootstrap

Parametric g-formula for point-exposures

One downside of the g-formula is that bootstrapping is often the only way to get confidence intervals.[12] Bootstrapping means we resample from the data and steps 1-3 each sample. Then we take the distribution of the bootstrap estimates as our measure of variation (e.g. variance/standard error).

This is easy in R, using the 'boot' function.

```
bootsamples = boot(data=coalplant, statistic=gformula_meandiffs, R=500)
```
13

But SAS requires custom coding, usually via Macros

[12] This is computationally intensive

[13] The gformula_meandiffs function performs steps 1-3, see code

Example 1: results

Parametric g-formula for point-exposures

Effect measures (mean differences):

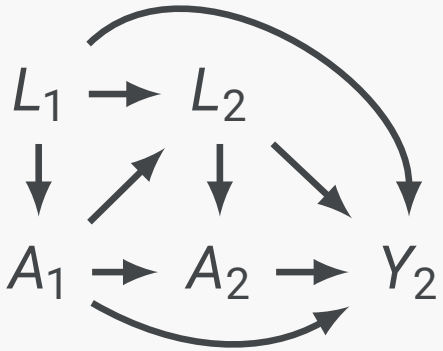
G-formula: policy estimates. Mean MDI, MDI difference (compared to natural course) and bootstrap confidence intervals	
Intervention/Effect	Mean/mean difference (95% CI)
a) Mean MDI: Natural course	97.32 (96.81, 97.82)
b) Effect: Shutdown	2.668 (1.182, 4.154)
c) Effect: Attr. MDI Cd	1.253 (0.449, 2.057)
d) Effect: Attr. MDI As	-.234 (-1.66, 1.191)
e) Effect: Attr. MDI Be	1.842 (0.687, 2.997)

Interpretation: decommissioning the coal plant for one year would result in an increase of 2.7 points in the mean MDI of 2 year olds, relative to doing nothing. This joint effect results in a larger difference that completely eliminating any exposure on its own.

The parametric g-formula for a point exposure involves only a typical regression model, but yields vast improvement in inference - both in terms of yielding more useful effect measures than the model default parameters (e.g. odds ratios), and in terms of the ease with which we can characterize effects of joint exposures.

There is a stark contrast in the simplicity of implementing the g-formula with a point exposure, versus with longitudinal data.

The g-formula: two times[14]



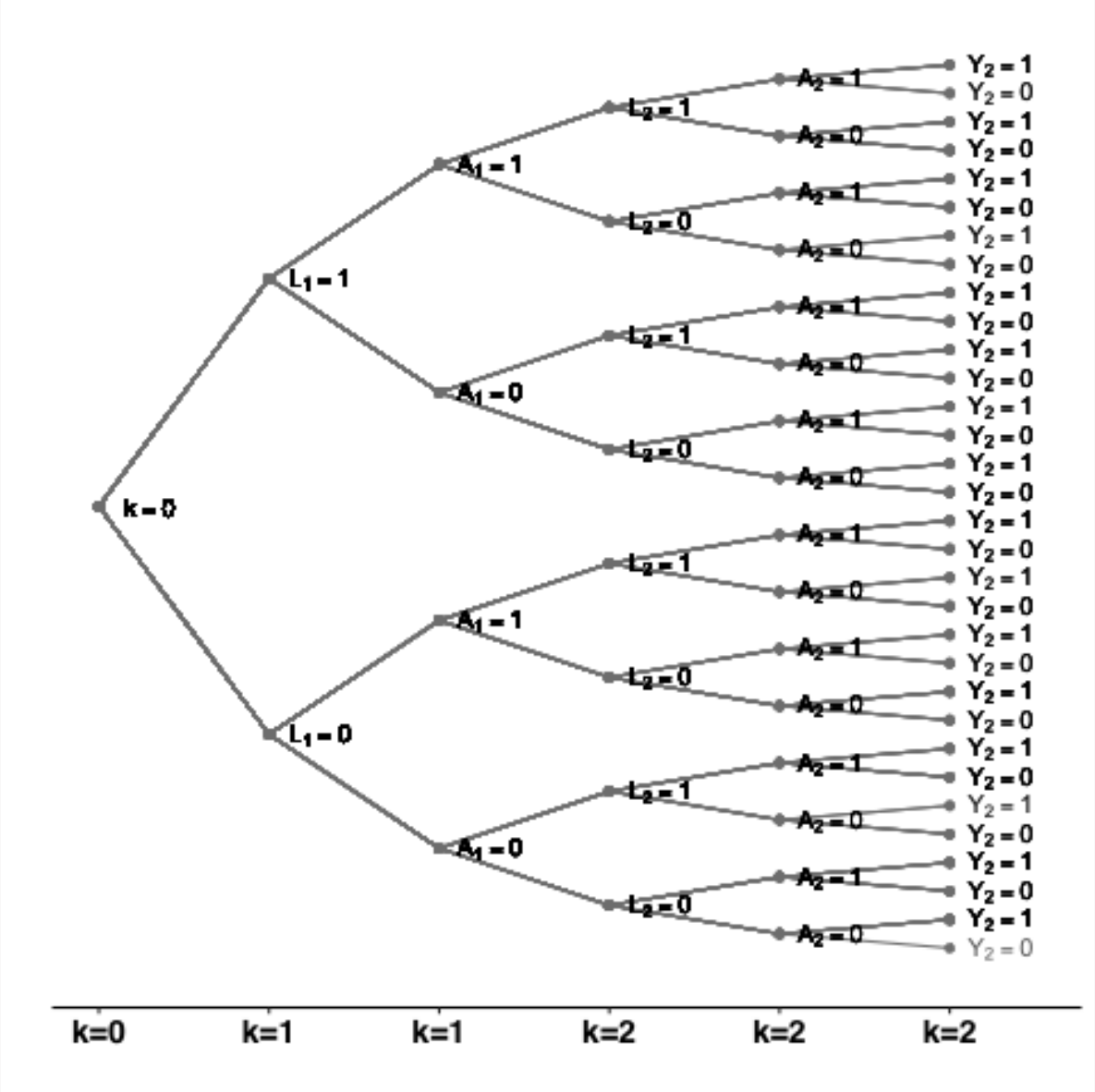
Evaluate at $A_1 = A_2 = g$

$$Pr\left(Y_2^g = 1\right) =$$
$$\sum_{\ell_1}\sum_{\ell_2} Pr\left(Y_2 = 1 | g, \ell_1, \ell_2\right) p\left(g, \ell_1, \ell_2\right)$$
$$= \sum_{\bar{\ell}_2} Pr\left(Y_2 = 1 | g, \bar{\ell}_2\right) p\left(\ell_2 | g, \ell_1\right) p\left(\ell_1\right)$$

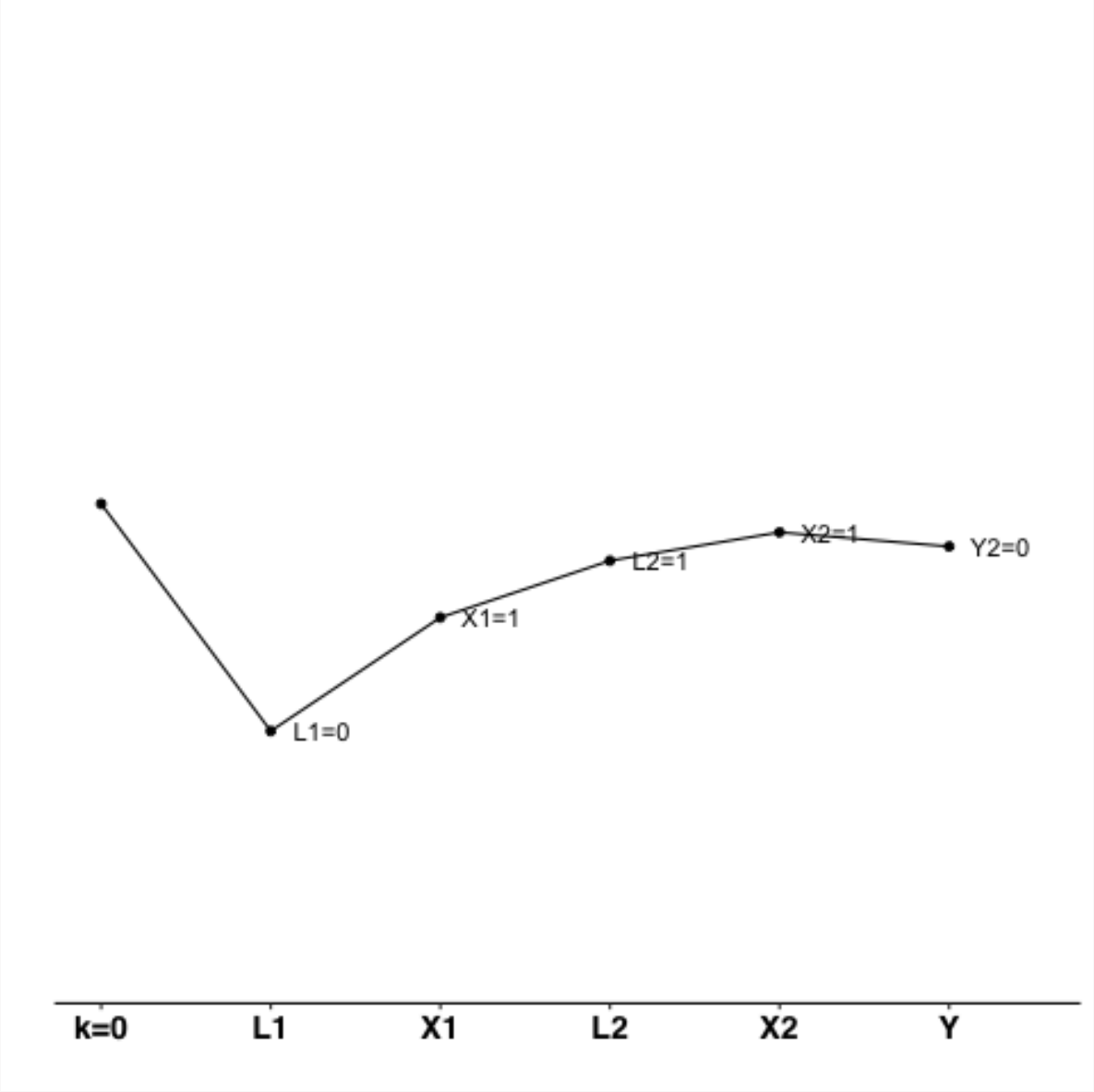
[14] sometimes called the longitudinal g-formula

We can represent these probabilities as nodes on a tree diagram

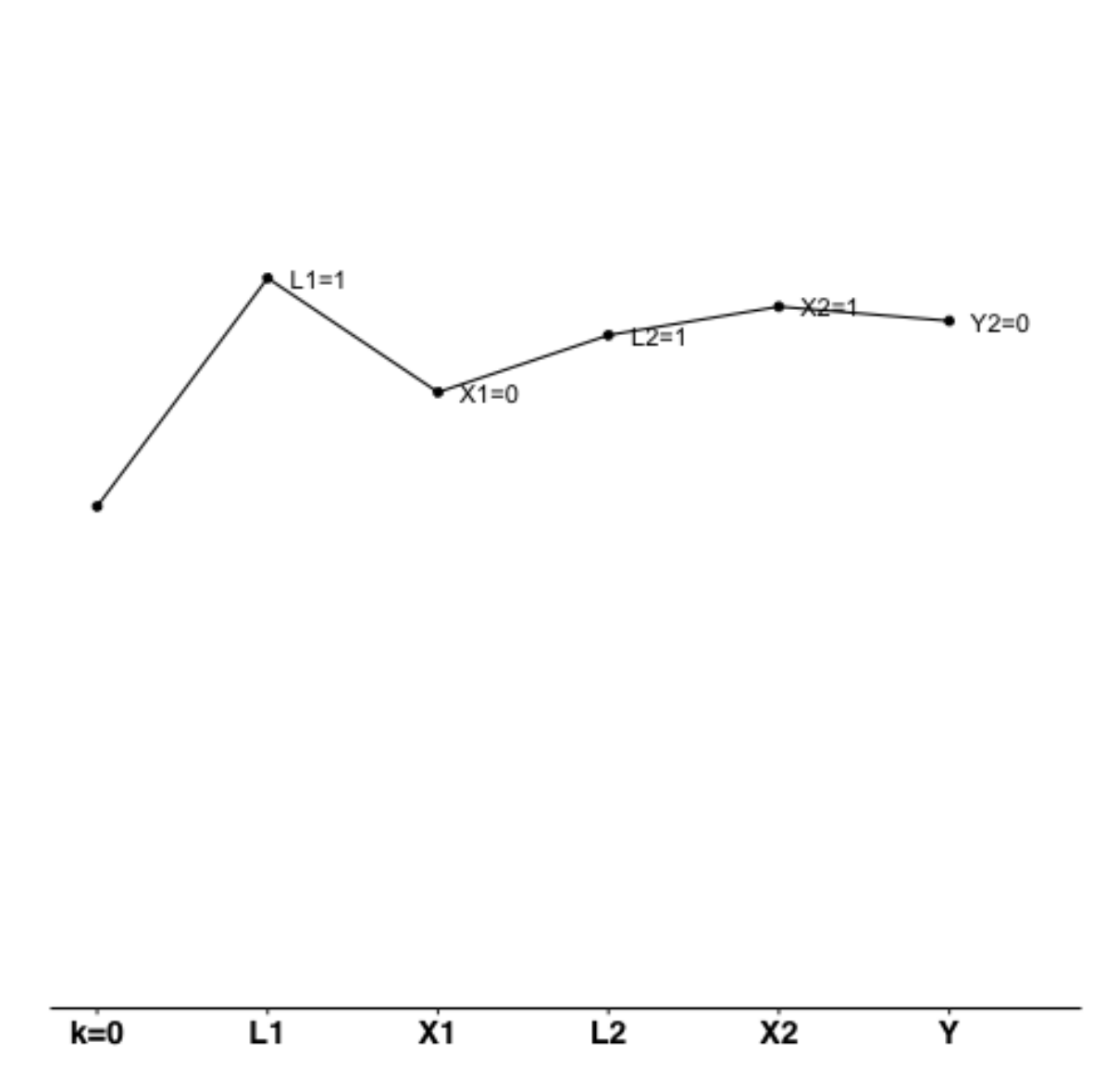
What are the possibilities (i.e. what are possible values of (y_2, a_k, l_k))?



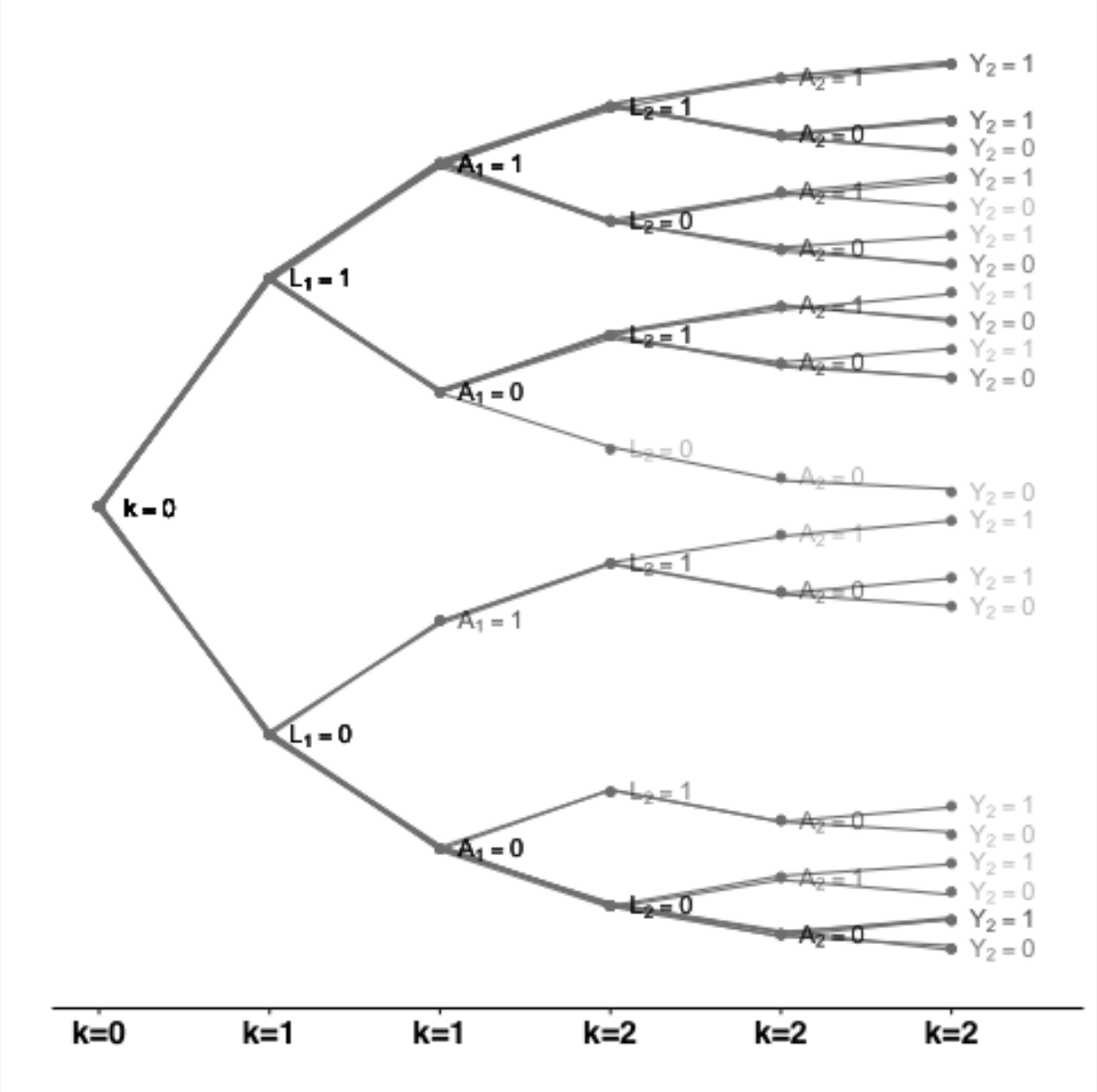
One possible realization



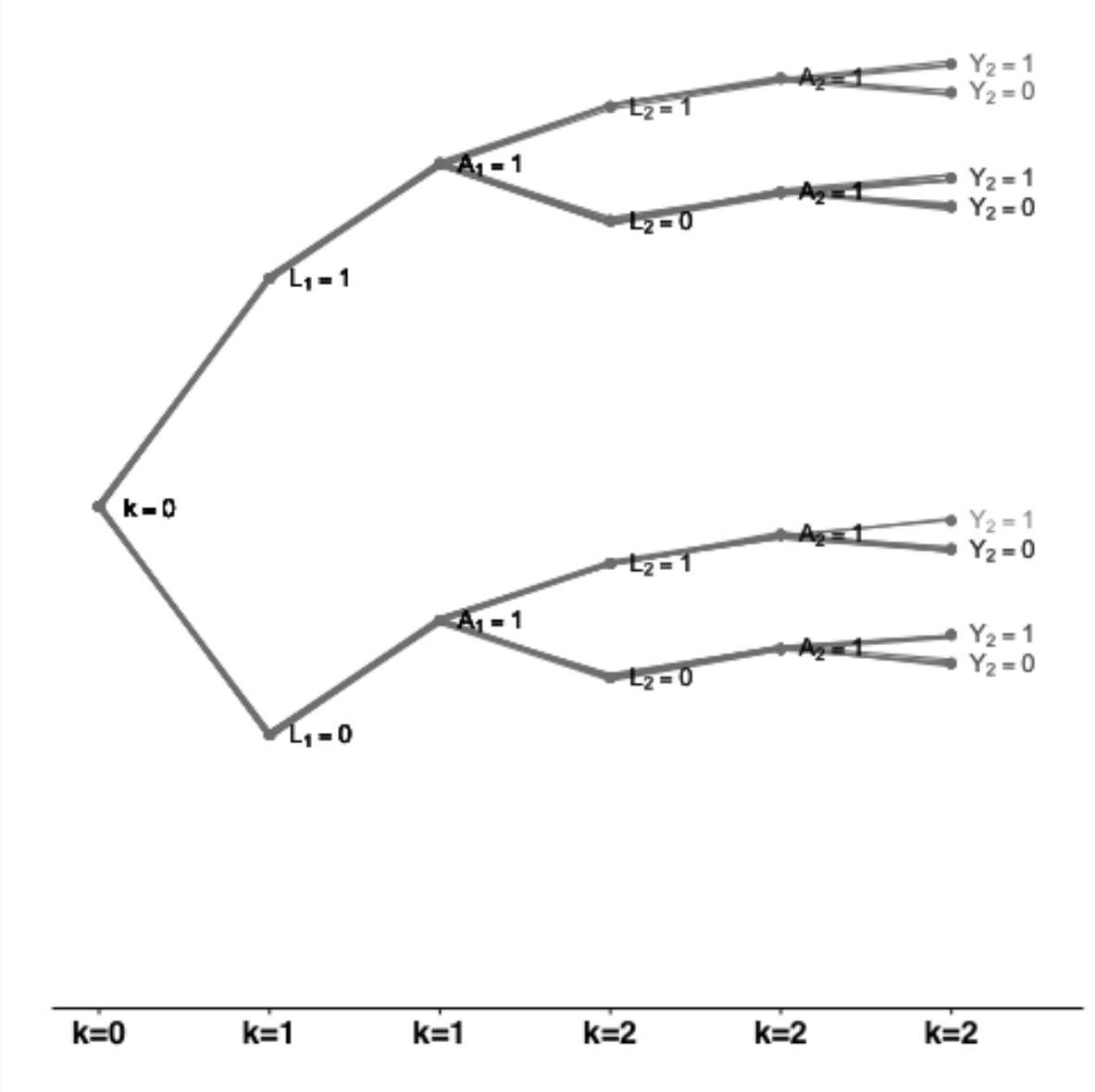
Another possible realization



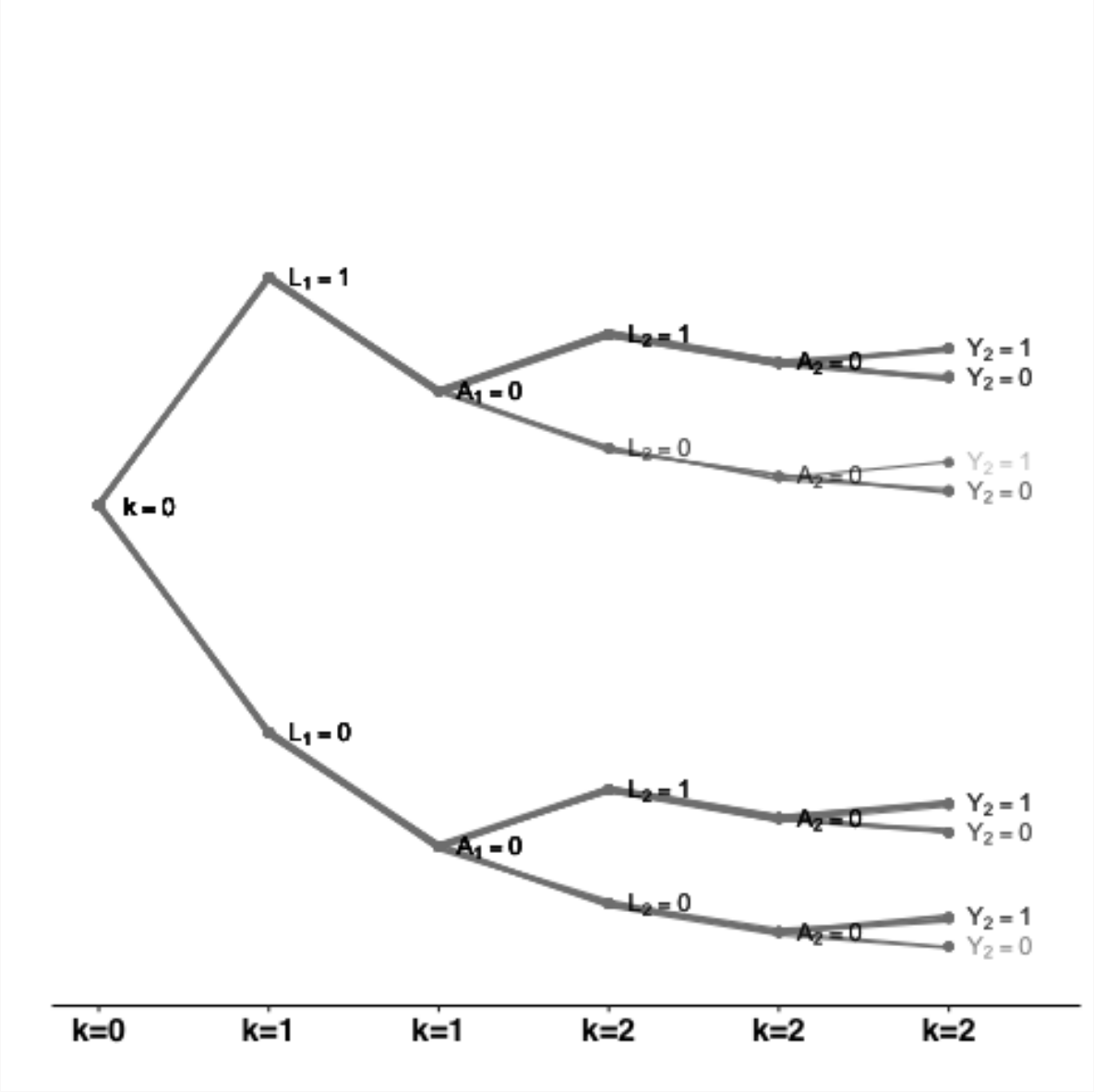
Observed population or population under the regime ‘Natural course’



A population under the regime ‘Always exposed’



A population under the regime ‘Never exposed’



Towards estimation

Intuition from tree diagrams

We want

$$Pr\left(Y_2^g = 1\right) = \sum_{\bar{\ell}_2} Pr\left(Y_2 = 1 | g, \bar{\ell}_2\right) p\left(\ell_2 | g, \ell_1\right) p\left(\ell_1\right)$$

Rather than enumerating every probability on the diagram and doing this math, we can simulate individuals through time

Why? $Pr\left(Y_2^g = 1\right)$ can be read directly off the diagrams where we intervened, via sample averages, so we can just simulate the diagrams!

Monte Carlo algorithm

Estimating $E(Y_2^g)$ or $p(y_2^g)$ without doing much math

1. Sample, with replacement, from the target population at time $k = 0$, $\widehat{p}(\ell_1) = \widehat{p}_n(\ell_1)$ (**pseudo-population** N^g)
2. Set A_1 equal to g for everyone in the pseudo-population
3. Simulate values of L_2 from $p\left(\ell_2 \middle| g, \ell_1\right)$ (e.g. a Bernoulli distribution with $\mu = p\left(\ell_2 \middle| g, \ell_1\right)$)
4. Simulate values of Y_2 from $p\left(Y_2 \middle| g, \ell_1, \ell_2\right)$
5. $E(Y^g)$ is just the mean of Y_2 in the simulated pseudo-population

Monte Carlo algorithm

Moving beyond $k = 2$, binary covariates

$(\bar{a}_k, \bar{\ell}_k, \bar{y}_k)$ **possibilities grow faster than data:** For $k = 2$, with two binary, time-varying variables and one outcome, there were already $2^5 = 32$ values of $(\bar{a}_k, \bar{\ell}_k, \bar{y}_k)$

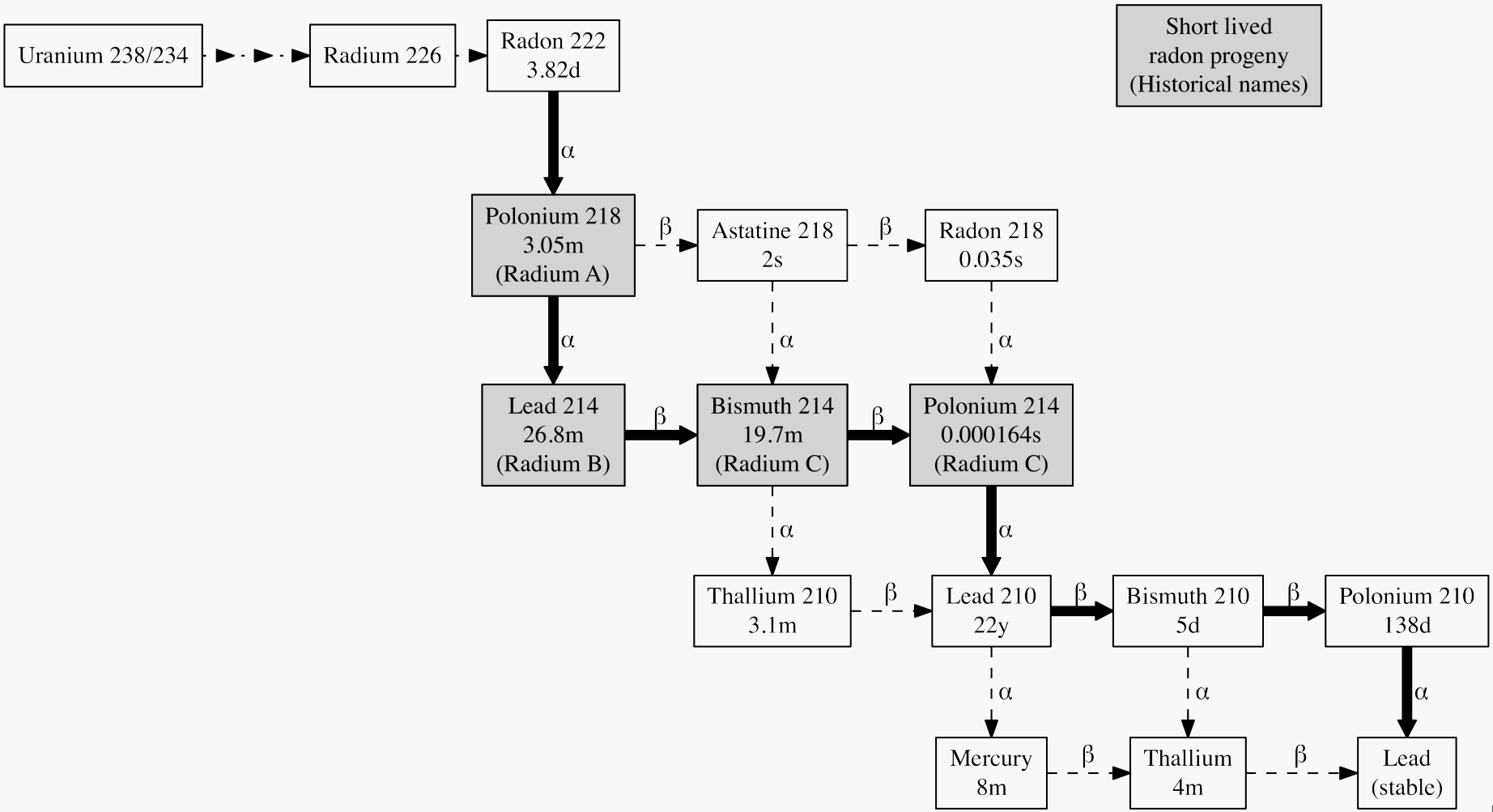
Data hungry: In general, we need estimates of $p\left(\ell_k | \bar{a}_{k-1}, \bar{\ell}_{k-1}, \bar{y}_{k-1}\right), p\left(y_k | \bar{a}_{k-1}, \bar{\ell}_{k-1}, \bar{y}_{k-1}\right)$,[15] which we won't be able to estimate non-parametrically (i.e. with sample proportions)

We have to assume some stronger assumptions to make headway. Hence, we need to use models as in example 1, even if we don't have continuous covariates.

[15] For dynamic regimes, we may also need $p\left(a_k | \bar{a}_{k-1}, \bar{\ell}_{k-1}, \bar{y}_{k-1}\right)$

Examples 2&3: radon and mortality

Analysis of occupational standards and dose-response



Examples 2&3: radon and mortality

Analysis of occupational standards and dose-response

ORIGINAL ARTICLE

Occupational Radon Exposure and Lung Cancer Mortality

Estimating Intervention Effects Using the Parametric g-Formula

Jessie K. Edwards,[a] Leah J. McGrath,[a] Jessie P. Buckley,[a] Mary K. Schubauer-Berigan,[b] Stephen R. Cole,[a] and David B. Richardson[a]

Examples 2&3: radon and mortality

Analysis of occupational standards and dose-response

ORIGINAL ARTICLE

Occup

Estin

Je

American Journal of Epidemiology

DOI: 10.1093/aje/k

Practice of Epidemiology

Healthy Worker Survivor Bias in the Colorado Plateau Uranium Miners Cohort

Alexander P. Keil*, David B. Richardson, and Melissa A. Troester

* Correspondence to Dr. Alexander P. Keil, Department of Epidemiology, Gillings School of Global Public Health, University of North Caro 2102E McGavran-Greenberg Hall, Campus Box 7435, Chapel Hill, NC 27599-7435 (e-mail: akeil@unc.edu).

Initially submitted July 13, 2014; accepted for publication November 17, 2014.

Examples 2&3: radon and mortality

Analysis of occupational standards and dose-response

ORIGINAL ARTICLE

American Journal of Epidemiology

ORIGINAL ARTICLE

Hypothetical Interventions to Limit Metalworking Fluid Exposures and Their Effects on COPD Mortality

G-Estimation Within a Public Health Framework

Sally Picciotto,[a] Jonathan Chevrier,[a,b] John Balmes,[a,c] and Ellen A. Eisen[a,b]

2102E McGavran-Greenberg Hall, Campus Box 7435, Chapel Hill, NC 27599-7435 (e-mail: akeil@unc.edu).

Initially submitted July 13, 2014; accepted for publication November 17, 2014.

Example 2: methods

Parametric g-formula

- Target population: the study population of uranium miners
- Estimate risk (cumulative incidence) from years 0-20 on the study we would have observed under:
  - No intervention (natural course)
  - **0.3** $\times$ 1000 pCi/L-year limit (slightly below current standard)
  - **0.1** $\times$ 1000 pCi/L-year limit (even lower standard)
  - No exposure (attributable risk)
- The new (hypothetical) standards are dynamic regimes: "If at work, then be exposed below the limit, and unexposed otherwise."

miners.csv

First 10 observations

id	intime	outtin	dead	rad	rad_lag1	cum_rad	cum_ra	atwork	leavework	durwork_	durwork	smoker
1	0.00	1.00	0	0.033	0.000	0.033	0.000	1	0	0	1	0
1	1.00	2.00	0	0.000	0.033	0.033	0.033	0	1	1	1	0
1	2.00	3.00	0	0.000	0.000	0.033	0.033	0	0	1	1	0
1	3.00	4.00	0	0.000	0.000	0.033	0.033	0	0	1	1	0
1	4.00	4.21	1	0.000	0.000	0.033	0.033	0	0	1	1	0
2	0.00	0.48	1	0.141	0.000	0.141	0.000	1	0	0	1	0
3	0.00	1.00	0	0.111	0.000	0.111	0.000	1	0	0	1	1
3	1.00	2.00	0	0.012	0.111	0.123	0.111	1	0	1	2	1
3	2.00	2.18	1	0.000	0.012	0.123	0.123	0	1	2	2	1
4	0.00	1.00	0	0.130	0.000	0.130	0.000	1	0	0	1	0

Note the multiple observations per individual (each record represents a person year)

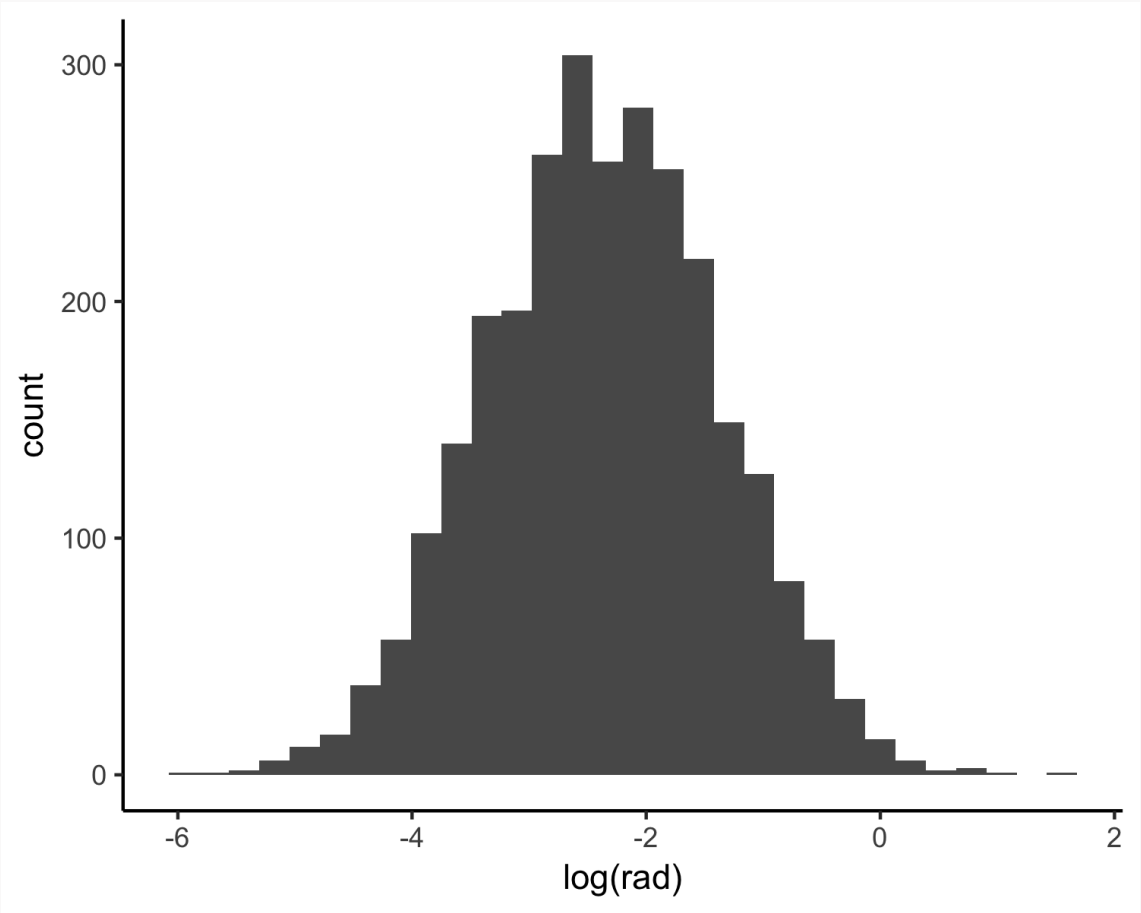
Example 2: methods

Parametric g-formula, Monte Carlo algorithm

1. Fit models to the observed data $p(a_k|\cdot), p(l_k|\cdot), p(y_k|\cdot)$
2. Take a large sample from the data at baseline (sample average estimator of baseline confounder distribution $p(l_0)$)
3. Simulate from the models in step 1 as well as the user specified intervention distribution $p_d(a_k|\cdot)$ (e.g. $a_k = 0$ at all times)
4. Effect estimation using the simulated outcomes (which represent Y^g)

Step 1: modeling

Parametric g-formula to estimate intervention effects



[15]Modeling choices: linear model (for continuous variables) and logistic models (for binary variables) are common choices of models for fitting the observed data. Here, annual exposure looks roughly log-normal, so we will model the log of the 'rad' variable. Unlike standard regression models, taking the log of exposure does not change the interpretation of the results!

Step 1: modeling, continued

Parametric g-formula to estimate intervention effects

Model exposure (annual log-radon exposure)[16]

```
PROC GENMOD DATA = &indata;
  TITLE 'Exposure model for g-formula estimates of occupational-radon policy effects on mortality';
  WHERE atwork=1;
  MODEL lograd = outtime cum_rad_lag1 smoker / D=NORMAL LINK=ID;
  ODS OUTPUT ParameterEstimates = ecoefs (KEEP=parameter estimate);
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3349	0.0325	-2.3987	-2.2712	5152.72	<.0001
outtime	1	-0.0163	0.0124	-0.0407	0.0080	1.73	0.1883
cum_rad_lag1	1	0.1046	0.0747	-0.0418	0.2510	1.96	0.1613
smoker	1	-0.3564	0.0844	-0.5219	-0.1910	17.84	<.0001
Scale	1	0.9922	0.0132	0.9667	1.0184		

[16] Exposure is only modeled during employed person time. There will be differences between SAS and R coefficients at ~12th decimal place.

Step 1: modeling, continued

Parametric g-formula to estimate intervention effects

Model time-varying confounder(s) (employment)[16]

```
PROC GENMOD DATA = &indata DESCENDING;
  TITLE 'Employment (confounder) model for g-formula estimates of occupational-radon policy effects on mortality';
  WHERE atwork=1 or leavework=1;
  MODEL leavework = outtime outtime*outtime rad_lag1 cum_rad_lag1 smoker  / D=B LINK=LOGIT;
  ODS OUTPUT ParameterEstimates = wcoefs (KEEP=parameter estimate WHERE=(parameter ^= "Scale"));
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.9626	0.2344	-4.4220	-3.5032	285.81	<.0001
outtime	1	0.7950	0.1205	0.5589	1.0312	43.54	<.0001
outtime*outtime	1	-0.0592	0.0125	-0.0836	-0.0348	22.59	<.0001
rad_lag1	1	0.8513	0.6457	-0.4144	2.1169	1.74	0.1874
cum_rad_lag1	1	-1.3723	0.4234	-2.2021	-0.5424	10.50	0.0012
smoker	1	1.6203	0.2150	1.1988	2.0417	56.77	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

[16] We model 'leaving' employment, since in these data once a miner leaves he/she does not return. This is among the person time that includes the time intervals at work and the time-interval in which the miner leaves work.

Step 1: modeling, continued

Parametric g-formula to estimate intervention effects

Model time-varying outcomes(s) (death)

```
PROC GENMOD DATA = &indata DESCENDING;
  TITLE 'Death (outcome) model for g-formula estimates of occupational-radon policy effects on mortality';
  MODEL dead = outtime atwork cum_rad cum_rad*cum_rad smoker smoker*cum_rad / D=B LINK=LOGIT;
  ODS OUTPUT ParameterEstimates = dcoefs (KEEP=parameter estimate WHERE=(parameter ^= "Scale"));
RUN;
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5835	0.1277	-0.8337	-0.3332	20.89	<.0001
outtime	1	-0.1480	0.0200	-0.1872	-0.1088	54.66	<.0001
atwork	1	-1.2275	0.1173	-1.4574	-0.9976	109.49	<.0001
cum_rad	1	1.6997	0.2509	1.2079	2.1914	45.89	<.0001
cum_rad*cum_rad	1	-0.1443	0.1147	-0.3691	0.0806	1.58	0.2086
smoker	1	1.4562	0.1999	1.0645	1.8479	53.09	<.0001
cum_rad*smoker	1	-1.7758	0.7646	-3.2744	-0.2773	5.39	0.0202
Scale	0	1.0000	0.0000	1.0000	1.0000		

Step 2: Sampling

Parametric g-formula to estimate intervention effects

Take a large sample from the data at baseline (Here we sample 50,000 miners into the pseudo-cohort)

In sas:

```
PROC SURVEYSELECT DATA=miners(WHERE=(intime=0))
   OUT=pseudo_cohort_bl OUTHITS N=&mc_iter METHOD=URS NOPRINT;
```

In R:

```
# large sample from observed baseline data
mc_iter = 50000
set.seed(12325)
mcdata <- slice(baseline_data, sample(1:N, size = mc_iter, replace=TRUE))
```

Step 3: Simulating

Parametric g-formula to estimate intervention effects

Simulate forward in time, emulating the assumed causal structure of the data[16]

In sas (nested DO loops):

```
%MACRO gformula_risks(indata, outdata, endtime=20, mc_iter=50000, seed=12312);
```

In R (as few loops as possible):

```
gformula_risks <- function(intervention=NULL, pseudo_cohort_bl,
                           endtime, mod_e, mod_w, mod_d, seed=NULL){
```

[16] I create custom R functions and custom SAS macros to do this step. In general, the available software for performing analysis with g-formula does not allow the types of interventions specified here, with at least one exception.

Step 3: Simulating, continued

Parametric g-formula to estimate intervention effects

E.g. Simulating employment over time [17]

In sas (nested DO loops):[18]

```
DO intime = 0 TO (&ENDTIME-1);
  outtime=intime+1;
  IF atwork THEN DO;
    IF intime>0 THEN DO;
      *in observed data, no one leaves in first year - enforce that here;
      *log-odds of leaving work;
      mul = _w[1] + _w[2]*outtime + _w[3]*outtime*outtime + _w[4]*rad_lag1 + _w[5]*cum_rad_lag1 + _w[6]*smoker;
      *random draw from a bernoulli distribution with probability of leaving work given
       by the inverse logit transform of the log-odds of leaving work;
      leavework = RAND('bernoulli', 1/(1+exp(-mul)));
    END;
    IF leavework = 1 THEN atwork = 0;
    durwork = durwork+atwork;
  END; *atwork;
```

[17] Note that the simulation of employment status is based on the model coefficients from step 1, and previous simulated versions of each variable

[18] DATA steps in SAS work by looping over observations in the dataset, so we simulate forward in time for each individual in the pseudo-cohort

Step 3: Simulating, continued

Parametric g-formula to estimate intervention effects

E.g. Simulating employment over time [17]

In R (as few loops as possible):[18]

```
for(t in seq(1, endtime, 1)){
  # index that keeps track of time
  idx = which(pseudo_cohort$outtime == t)
  ############
  # leaving work (time varying confounder)
  ############
  if(t==1) widx = 1:length(idx) # everyone is at work at first time point
  if(t > 1){
     widx = which(pseudo_cohort[idx-1,'atwork']==1)
     # index keeping track of which workers still work
     pseudo_cohort[idx[widx],'leavework'] <- rbinom(n=length(widx), size=1,
                                                    prob=predict(mod_w, newdata = pseudo_cohort[idx[widx],],
                                                                 type = 'response'))
     # if worker didn't leave, then assume stay at work
     pseudo_cohort[idx,'atwork'] <- (pseudo_cohort[idx,'leavework']==0 & pseudo_cohort[idx-1,'atwork']==1)
     # update at work index to account for workers who left
     widx = which(pseudo_cohort[idx,'atwork']==1)
     pseudo_cohort[idx,'durwork'] <- pseudo_cohort[idx-1,'durwork'] + pseudo_cohort[idx,'atwork']
     pseudo_cohort[idx,'durwork_lag1'] <- pseudo_cohort[idx-1,'durwork']
```

[17] Note that the simulation of employment status is based on the model coefficients from step 1, and previous simulated versions of each variable

[18] R is faster when we operate on vectors, so we use matrix indices to simulate forward in time for all individuals in the pseudo-cohort at the same time

Step 3: Simulating, continued

Parametric g-formula to estimate intervention effects

Intervention distributions; Static intervention: never exposed

Intervening with IF statements

In sas:

```
/*********
* static intervention: always unexposed
**********/
IF intervention = 0 THEN rad = 0;
```

In R:

```
if(is.numeric(intervention)){
  # static, deterministic intervention, e.g. set exposure = 0
  pseudo_cohort[idx,'rad'] <- intervention
```

Step 3: Simulating, continued

Parametric g-formula to estimate intervention effects

Intervention distributions; Dynamic intervention: natural course[19]

Intervening with random draws from a log-normal distribution

In sas:

```
mue = _e[1] + _e[2]*outtime  + _e[3]*cum_rad_lag1  + _e[4]*smoker ;
*log exposure is a random draw from the regression mean and std. dev. of the residuals, capped at empirical maximum;
DRAWLOGEXPOSURE: lograd = MIN(RAND('NORMAL', mue, _e[5]), LOG(7.321276));
rad = EXP(lograd);
```

In R:

```
# exposure is assumed log-normally distributed (=predicted value plus draw from the residuals)
meanlogr = predict(mod_e, newdata = pseudo_cohort[idx[widx],])
logr = meanlogr + eps_e[idx[widx]]
pseudo_cohort[idx[widx],'rad'] <- exp(pmin(log(max_e), logr))
```

[19] It's an "intervention" because we can only approximate the observed exposure distribution, so the natural course really represents an intervention to make exposure appear more like a known probability distribution.

Step 3: Simulating, continued

Parametric g-formula to estimate intervention effects

Intervention distributions; Dynamic intervention: hypothetical occupational limits [20]

In sas:

```
DRAWLOGEXPOSURE: lograd = MIN(RAND('NORMAL', mue, _e[5]), LOG(7.321276));
rad = EXP(lograd);
*this antiquated bit of programming uses 'goto' statements to take another
* this is one way to draw from a truncated normal distribution with an upp
IF intervention = 1 AND rad>0.1 THEN GOTO DRAWLOGEXPOSURE;
ELSE IF intervention = 3 AND rad>0.3 THEN GOTO DRAWLOGEXPOSURE;
```

In R:

```
viol_idx <- with(pseudo_cohort[idx,], which(eval(parse(text=intervention))))
while(length(viol_idx)>0){
  # accept/rejection sampling to get draws from truncated log-normal
  # this is how we (I) assume an intervention would be implemented, but
  # we could imagine other ways (e.g. a hard cap on exposure)
  meanlogr = predict(mod_e, newdata = pseudo_cohort[idx[viol_idx],])
  logr = meanlogr + rnorm(n=length(viol_idx), mean=0, sd=sqrt(var_e))
  pseudo_cohort[idx[viol_idx],'rad'] <- exp(logr)
  viol_idx <- with(pseudo_cohort[idx,], which(eval(parse(text=intervention))))
  # check whether any are still in violation of the distribution, if so, repeat loop
}}} # end dynamic intervention
```

[20] We operationalize these by simulating from a truncated log-normal distribution, where the distribution has an upper truncation point at the new limit

Step 3: Simulating, continued

Parametric g-formula to estimate intervention effects

Simulating individual risk of death: NOT simulating deaths [21]

In sas:

```
/*********
* discrete hazard of death
**********/
mud = _d[1] + _d[2]*outtime + _d[3]*atwork + _d[4]*cum_rad + _d[5]*cum_rad*cum_rad + _d[6]*smoker + _d[7]*smoker*cum_rad;
pdead = 1/(1+exp(-mud));
*update cumulative incidence with kaplain meier estimator;
IF intime = 0 THEN cum_incidence = pdead;
ELSE cum_incidence = cum_incidence + (1-cum_incidence)*pdead;
```

In R:

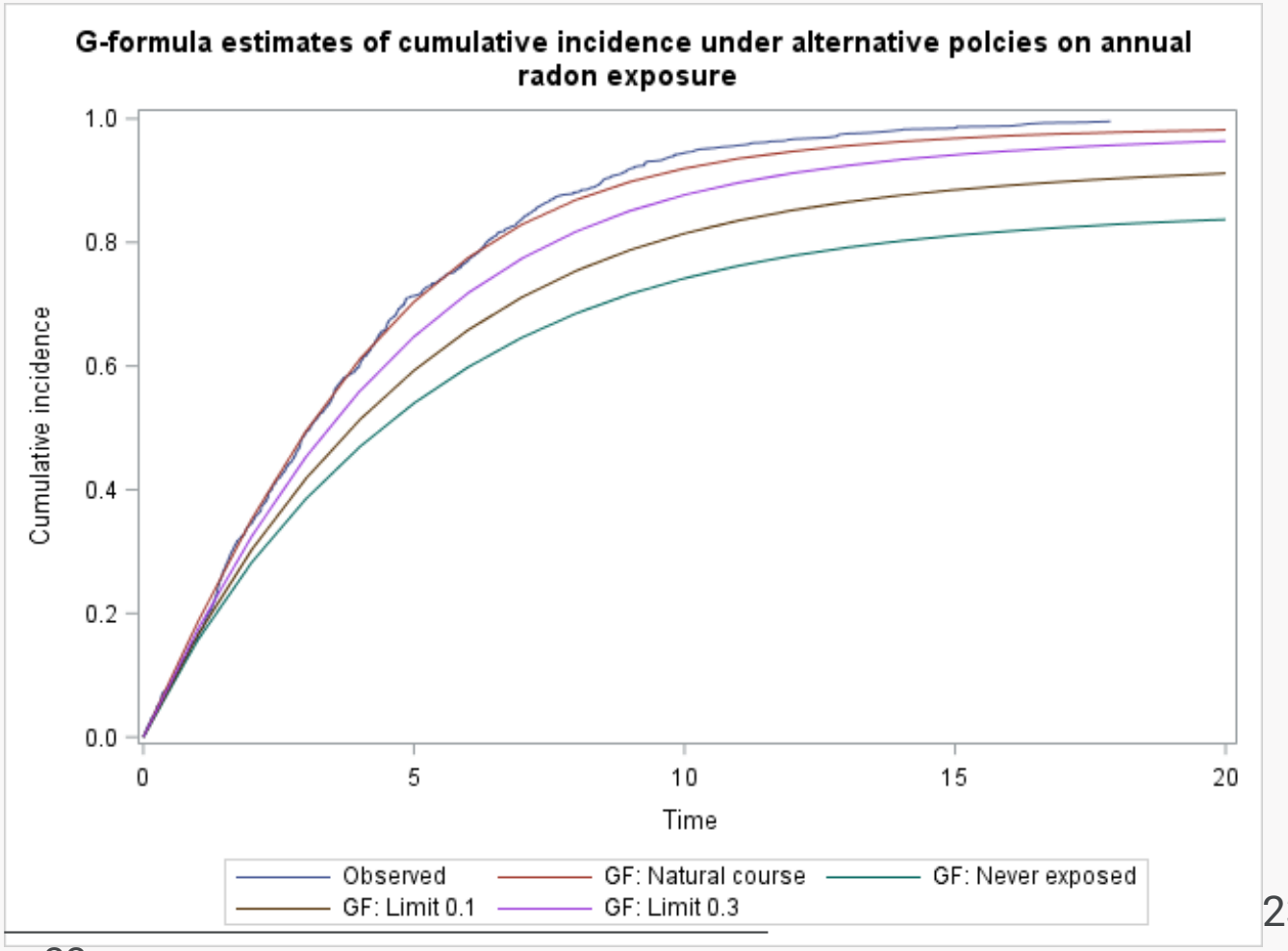
```
pseudo_cohort[idx,'dead'] = predict(mod_d, newdata = pseudo_cohort[idx,], type = 'response')
if(t > 1){
  # kaplain-meier estimator of cumulative incidence (note this is applied on the individual basis)
  pseudo_cohort[idx,'cum_incidence'] = pseudo_cohort[idx-1,'cum_incidence'] +
    (1-pseudo_cohort[idx-1,'cum_incidence'])*pseudo_cohort[idx,'dead']
} else{
  pseudo_cohort[idx,'cum_incidence'] = pseudo_cohort[idx,'dead']
}
```

[21] This relies on the fact that the average risk is equal to the average of the individual risks; we may wish to resolve to 1/0 if estimating hazard ratios, e.g.: D. Westreich et al.(2012) *Stat Med*;A. P. Keil et al.(2014) *Epidemiology*

Step 3: Simulating, continued

Parametric g-formula to estimate intervention effects

Cumulative incidence point estimatesin SAS[22]



[23]

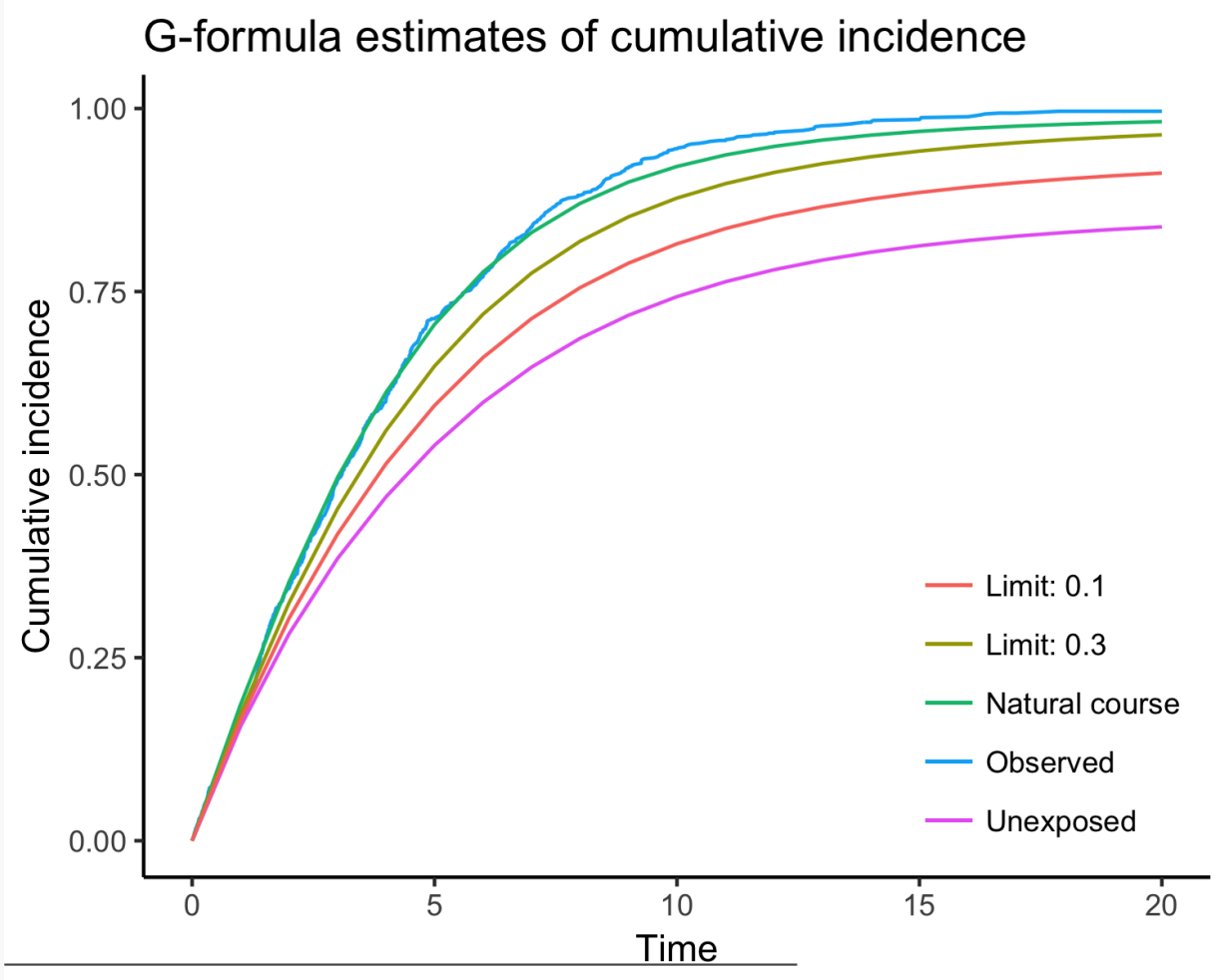
[22] These should be roughly the same in SAS and R, but we are subject to simulation error if the MC sample is not big enough

[23] SAS

Step 3: Simulating, continued

Parametric g-formula to estimate intervention effects

Cumulative incidence point estimatesin R[22]



[23]

[22] These should be roughly the same in SAS and R, but we are subject to simulation error if the MC sample is not big enough

[23] R

Step 4: Effect estimates

Parametric g-formula to estimate intervention effects

Survival time ratio: $E(T^{NC})/E(T^{Rad<0.1})$[24, 25]

In sas:

```
PROC SQL;
  TITLE 'Policy time ratio for comparison with structural nested model ';
  CREATE TABLE a AS
     SELECT
     SUM(1-cum_incidence) AS elifnc FROM pseudo_cohort(WHERE=(intervention=-1));
  CREATE TABLE b AS
     SELECT
     SUM(1-cum_incidence) AS elifexp1 FROM pseudo_cohort(WHERE=(intervention=1));
  SELECT elifexp1/elifnc AS TR FROM a, b;
QUIT;
```

Policy time ratio for comparison with structural nested model

TR
1.4744

In R:

```
> elifenc = with(nc, sum(1-cum_incidence))
> elifeexp1 = with(exp1, sum(1-cum_incidence))
> elifeexp1/elifenc
[1] 1.476379
```

[24] Really, we are estimating the restricted mean survival time ratio, which estimates the survival time ratio because most of the pseudo-cohort dies.

[25] Our structural nested AFT model will estimate this directly

Step 4: Effect estimates, cont.

Parametric g-formula to estimate intervention effects

Bootstrapping confidence intervals[26]

In sas:

```
%LET i = 1;
%DO %WHILE(%EVAL(&i<=&iter));
%PUT BOOTSTRAP_GF: iteration &i of &iter;
  PROC SURVEYSELECT DATA=&indata OUT=bootpop METHOD=URS N=&sampsize NOPRINT; SAMPLINGUNIT id;RUN;
  %get_coefs(bootpop);
  %gformula_risks(indata=bootpop, outdata=pseudo_cohort, endtime=&endtime, mc_iter=50000, seed=&i);
  %gformula_effectestimates(indata=pseudo_cohort, outdata=bootpe, endtime=&endtime);
  PROC APPEND DATA=bootpe(WHERE=(outtime=&endtime)) BASE=boot_samples;
%LET i = %EVAL(&I+1);
%END;
```

In R:

```
nbootsamples = 200
system.time(boot_samples <- boot(data = baseline_data, fdata=miners, statistic = gformula_effectestimates, R = nbootsamples, endtime=10))
```

[26] For bootstrapping, it helps to have the analysis done within a macro (SAS) or a function (R)

Step 4: Effect estimates, cont.

Parametric g-formula to estimate intervention effects

Bootstrapping confidence intervals for the 10 year risk difference: results[27]

In sas:

G-formula: policy estimates. 10-year Risk difference estimates (compared to natural course) and bootstrap confidence intervals	
Intervention	Risk difference (95% CI)
Limit = 0.3	-.043 (-.049,-.036)
Limit = 0.1	-.105 (-.124,-.086)
Never exposed	-.178 (-.216,-.140)

In R:

```
rd_exp3_10:-0.043 (-0.052, -0.034)
rd_exp1_10:-0.106 (-0.131, -0.082)
rd_unex_10:-0.180 (-0.227, -0.132)
```

[27] We could get more similar estimates between SAS and R by increasing the size of the MC sample, at the cost of computational time.

Step 4: Effect estimates, cont.

Parametric g-formula to estimate intervention effects

The g-formula doesn't give just one parameter - we get the full distribution of counterfactuals under any intervention we want (within reason). We estimated the time-ratio and the risk difference, but could easily get risk ratio, rate ratio/hazard ratio, or years of life lost. For example, see D. Westreich et al.(2012) *Stat Med*, A. P. Keil et al.(2014) *Epidemiology*, or A. P. Keil D. B. Richardson (2017) *Environ Health Perspect*.

Step 4: Effect estimates, cont.

Parametric g-formula to estimate intervention effects

The g-fo … get the full
distributi … ion we want
(within re … he risk
differenc … /hazard ratio,
or years … et al.(2012)
Stat Med … P. Keil D. B.
Richards …

TABLE 3. Hazard Ratios Comparing the Hazard of All-cause Mortality Between Patients With and Without Graft-versus-host Disease (Regression) or Comparing Cohorts With and Without Hypothetical Intervention to Prevent Graft-versus-host Disease (g-formula)

Method	HR	95% CI
Regression		
Crude	1.2	0.77–2.0
Baseline adjusted[a]	1.2	0.71–1.9
Fully adjusted[b]	2.3	1.4–3.9
g-formula		
Natural course vs. prevent[c]	1.1	0.91–1.3

[a]Baseline covariates include age at date of bone marrow transplant, wait time until transplant, sex, and cytomegalovirus status at baseline.

[b]Adjusted for baseline covariates above and time-varying covariates, including days during which platelets had not returned to normal, cumulative days the patient had not experienced relapse, and indicators for relapse and platelets returning to normal on a given day.

[c]Comparing the hazard of all-cause mortality between the entire cohort simulated under no intervention and the entire cohort of simulated to be unexposed (referent) at all time points.

Step 4: Effect estimates, cont.

Parametric g-formula to estimate intervention effects

TABLE 3. Hazard Ratios Comparing the Hazard of All-cause Mortality Between Patients With and Without Graft-versus-host Disease (Regression) or Comparing Cohorts With and Without Hypothetical Intervention to Prevent Graft-versus-host Disease (g-formula)

Table III. Hazard ratios and 95% confidence limits for always treated versus never treated with combination antiretroviral therapy.

Cox models	Hazard ratio	95% confidence limits	Point estimate (SE)
No covariates*	0.94	0.74, 1.19	−0.066 (0.123)
Baseline covariates, unweighted	0.66	0.50, 0.87	−0.414 (0.140)
Baseline and time-varying covariates, unweighted*	0.75	0.58, 0.96	−0.294 (0.129)
Baseline covariates, weighted*	0.56	0.42, 0.75	−0.575 (0.147)
Parametric g-formula†	0.55	0.42, 0.71	−0.606 (0.133)

SE, standard error

Baseline covariates were age, sex, weight, visit, (baseline) viral load, and CD4 count. Time-varying covariates were viral load and CD4 count.

*Results from the work of Cole *et al.* were 0.98 (95% confidence limits (CL): 0.76, 1.26), 0.81 (95% CL: 0.61, 1.07), and 0.54 (95% CL: 0.38, 0.78).

†Adjusted for all baseline and time-varying covariates.

under no intervention and the entire cohort of simulated to be unexposed (referent) at all time points.

Step 4: Effect estimates, cont.

Parametric g-formula to estimate intervention effects

many more years of follow-up did not (Lubin et al., 2000). As an alternative measure of arsenic's impact across the life course, the g-formula allowed us to calculate years of life lost for all causes by simply comparing the person-time under each intervention, which suggested an overall detriment that is not apparent in the cumulative incidence at age 90. This approach is not

TABLE 3. Hazard Ratios Comparing the Hazard of All-cause Mortality Between Patients With and Without Graft-versus-

Baseline covariates, weighted*	0.56	0.42, 0.75	−0.575 (0.147)
Parametric g-formula†	0.55	0.42, 0.71	−0.606 (0.133)

SE, standard error

Baseline covariates were age, sex, weight, visit, (baseline) viral load, and CD4 count. Time-varying covariates were viral load and CD4 count.

*Results from the work of Cole *et al.* were 0.98 (95% confidence limits (CL): 0.76, 1.26), 0.81 (95% CL: 0.61, 1.07), and 0.54 (95% CL: 0.38, 0.78).

†Adjusted for all baseline and time-varying covariates.

under no intervention and the entire cohort of simulated to be unexposed (referent) at all time points.

Longitudinal g-formula

Summarizing features

Flexibility: one can answer many types of questions about counterfactual distributions

Direct interpretation: parameters of interest are not restricted to model parameters

Often model heavy: high reliance on model accuracy for correct inference

Dose-response: comparisons to standard dose-response estimators is more difficult than answering questions about interventions

Longitudinal g-formula

Complications not discussed here

In order of difficulty:

Censoring, late entry, how to choose models

Structural nested models[28]

[28] Unofficial: designed to overcome the difficulty of estimating dose-response using the g-formula

Structural nested models (SNMs) parameterize ‘blip’ effects: the effect of a brief blip of exposure. These blip effects are conditional on being at the referent value of the exposure after the blip.
E.g. “what is the effect of exposure A=1 at time k, versus A=0, given that the population will be untreated from time $k + 1$ onwards”[29]

[29] Yes, this is weird

SNMs

From Robins 1999 (p 10):

"A SNM is model for the magnitude of the causal effect of a final brief blip of a time-dependent treatment at time t as a function of past time-dependent treatment and prognostic factor history... The essential difference between MSMs and SNMs is that SNMs model the magnitude of the effect of a treatment given at t as a function of the prognostic factor history up to t.[30] In contrast, MSMs model the causal effect of treatment given at t only as a function of baseline prognostic factors."

[30] i.e. an SNM can include interaction terms with time-varying covariates

SNMs

We will focus on structural nested accelerated failure time models, which have been the most commonly used in analysis of (especially) occupational studies

G-estimation of a SNAFTM

Structural nested accelerated failure time model (SNAFTM)

- Based on the “strong version” of the Cox and Oakes’ accelerated failure time model (Robins 1992)

G-estimation of a SNAFTM

Structural nested accelerated failure time model (SNAFTM)

- Based on the “strong version” of the Cox and Oakes’ accelerated failure time model (Robins 1992)
- A “blip” of exposure at time *k*, ages you at a rate = $exp(\psi)$ times your baseline rate of age for the duration of $[k, k+1)$

SNAFTMs

Structural nested accelerated failure time models[32]

One example: dose-response for cumulative exposure

$$T^0 = \int_0^T exp(\psi \bar{A}_k) dk$$

cumulative exposure $\bar{A}_k$

observed event time T

potential event time under no exposure T^0

"time ratio" $exp(\psi)$; find $\hat{\psi}$ with g-estimation[31]

---

[31] J. Chevrier S. Picciotto E. A. Eisen (2012) *Epidemiology*

[31] More completely: $T^{\bar{a}_m, \bar{0}} = m + \int_m^T exp(\psi \bar{A}_k) dk$

[32] M. A. Hernán et al.(2005) *Pharmacoepidemiol Drug Saf*

G-estimation

- Not the same as g-computation

G-estimation

- Not the same as g-computation
- Involves model for exposure $E(A_k|\bar{A}_{k-1}, \bar{L}_k)$ [like IPW]

G-estimation

- Not the same as g-computation
- Involves model for exposure $E(A_k|\bar{A}_{k-1}, \bar{L}_k)$ [like IPW]
- Iterative guesses at T^0

G-estimation

- Not the same as g-computation
- Involves model for exposure $E(A_k|\bar{A}_{k-1}, \bar{L}_k)$ [like IPW]
- Iterative guesses at T^0
- Repeat until $A_k \perp T^0|A_{k-1}, L_k$

G-estimation

- Not the same as g-computation
- Involves model for exposure $E(A_k|\bar{A}_{k-1}, \bar{L}_k)$ [like IPW]
- Iterative guesses at T^0
- Repeat until $A_k \perp T^0|A_{k-1}, L_k$
- CI found via "inverting a test statistic"

Basic g-estimation

1. Take a(n) (educated) guess at ψ, call the guess $\tilde{\psi}$

[33] In the code I use a more efficient version given by $\tilde{T}^{\bar{a}_m, \overline{0}} = m + \int_m^T exp(\tilde{\psi}\bar{A}_k)dk$

Basic g-estimation

1. Take a(n) (educated) guess at ψ, call the guess $\tilde{\psi}$
2. Use $\tilde{\psi}$ to generate a guess for T^0 (called $\tilde{T}_0$) from the data[33]

$$\tilde{T}_0 = \int_0^T \exp(A_k \tilde{\psi}) dk$$

[33] In the code I use a more efficient version given by $\tilde{T}^{\bar{a}_m, \bar{0}} = m + \int_m^T exp(\tilde{\psi} \bar{A}_k) dk$

Basic g-estimation

1. Take a(n) (educated) guess at ψ, call the guess $\tilde{\psi}$
2. Use $\tilde{\psi}$ to generate a guess for T^0 (called $\tilde{T}_0$) from the data[33]

$$\tilde{T}_0 = \int_0^T \exp(A_k \tilde{\psi}) dk$$

3. Include $\tilde{T}_0$ as baseline covariate in an exposure model (e.g.)

$$logit[Pr(A_k = 1 | \bar{L}_k, \bar{A}_{k-1}, V, \tilde{T}_0; \beta)] =$$

$$\beta_0 + \beta_1 L_k + \beta_2 L_{k-1} + \beta_3 A_{k-1} + \beta_4 V + \beta_5 \tilde{T}_0$$

[33] In the code I use a more efficient version given by $\tilde{T}^{\bar{a}_m, \bar{0}} = m + \int_m^T exp(\tilde{\psi} \bar{A}_k) dk$

Basic g-estimation

1. Take a(n) (educated) guess at ψ, call the guess $\tilde{\psi}$
2. Use $\tilde{\psi}$ to generate a guess for T^0 (called $\tilde{T_0}$) from the data[33]

$$\tilde{T_0} = \int_0^T \exp(A_k \tilde{\psi}) dk$$

3. Include $\tilde{T_0}$ as baseline covariate in an exposure model (e.g.)

$$logit[Pr(A_k = 1 | \bar{L}_k, \bar{A}_{k-1}, V, \tilde{T_0}; \beta)] =$$

$$\beta_0 + \beta_1 L_k + \beta_2 L_{k-1} + \beta_3 A_{k-1} + \beta_4 V + \beta_5 \tilde{T_0}$$

4. Repeat 1-3 until $\beta_5 = 0$ ($\tilde{\psi} = \psi$ if $\tilde{T_0} \perp A_k | \bar{A}_{k-1}, \bar{L}_k, V$)

[33] In the code I use a more efficient version given by $\tilde{T}^{\bar{a}_m, \bar{0}} = m + \int_m^T exp(\tilde{\psi} \bar{A}_k) dk$

Example 3: methods

g-estimation of a structural nested model

Step 1: make a bunch of guesses $\tilde{\psi}$; SAS:

```
DATA gestimation;
  SET miners;
  DO psi = 0 TO 1 BY 0.01;
```

R:

```
g_estimation <- function(psi.seq = seq(0, 1, 0.01)){
  # the main function
  chisq = numeric(length(psi.seq))
  for(i in 1:length(psi.seq)){
```

Example 3: methods

g-estimation of a structural nested model

Step 2: Calculuate $\tilde{T}^0$ (called 't_blip' here)

```
pyr0 = EXP(psi*cum_rad)*(outtime-intime);
```

```
DATA gestimation;
  SET gestimation;
  BY psi id DESCENDING intime;
  RETAIN revcumpyr0;
  IF first.id THEN revcumpyr0=0;
  revcumpyr0 = revcumpyr0 + pyr0;
  t_blip = revcumpyr0 + intime;
RUN;
```

Example 3: methods

g-estimation of a structural nested model

Step 3: Fit exposure model, including $\tilde{T}^0$ as a variable (called 't_blip' here), and calculate test statistic for the parameter for $\tilde{T}^0$.

```
PROC GENMOD DATA = gestimation2;
  TITLE 'Exposure model for g-estimation of log-time ratio per unit of cumulative radon exposure';
  BY psi;
  WHERE atwork=1;
  MODEL lograd = t_blip intime intime*intime cum_rad_lag1 smoker;
  ODS OUTPUT ParameterEstimates = gfunction(KEEP=chisq psi parameter WHERE=(parameter="t_blip"));
```

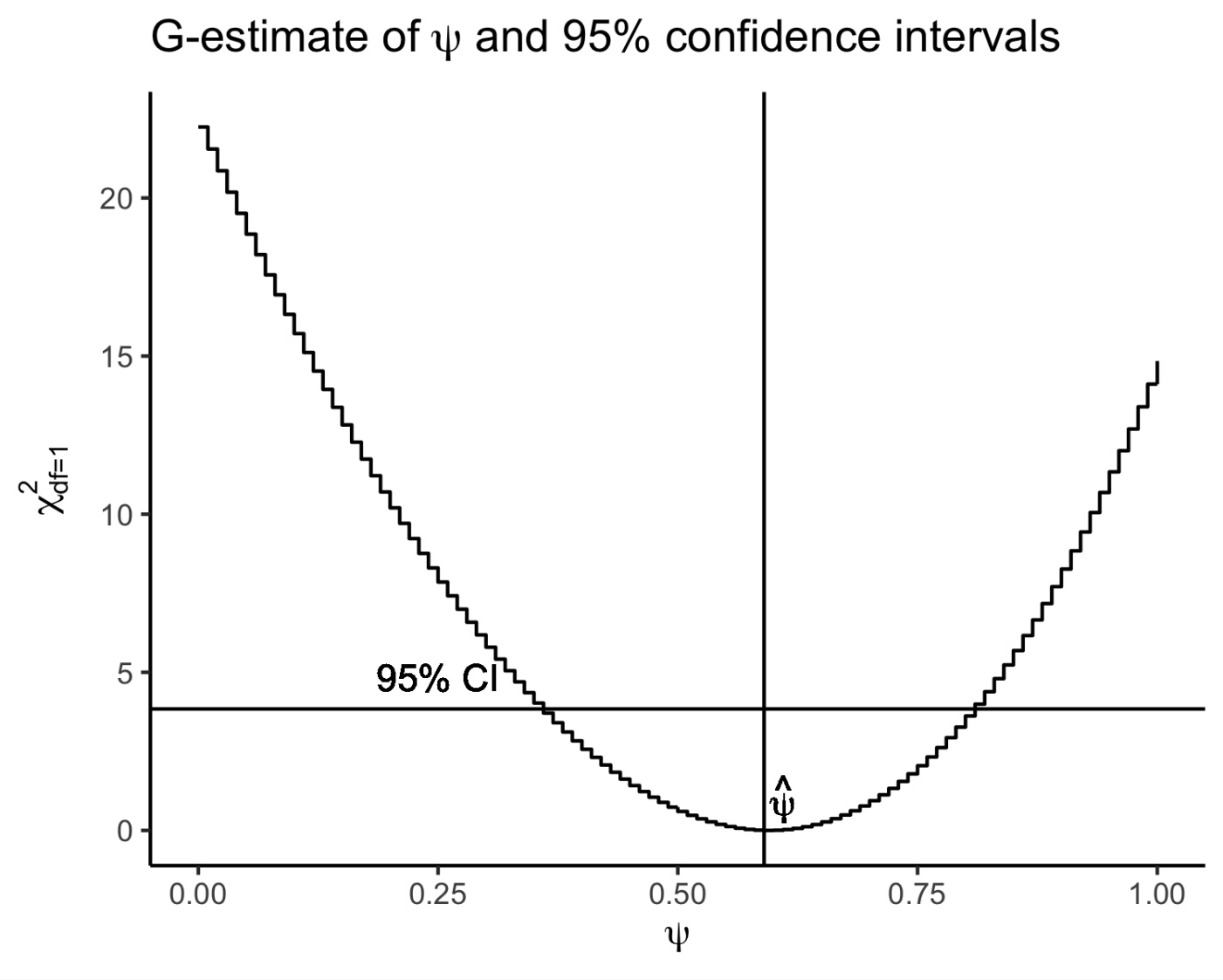
[34]

[34] Note that this model is fit only among the person time at work. This approach is taking advantage of 'selective ignorability' - this avoids the problem of non-positivity (e.g. exposure doesn't occur outside work). This is nearly always the approach taken in occupational data.

Example 3: results

g-estimation of a structural nested model

The ‘g-function’ and ‘inverting the test-statistic’ to get 95% CI



Example 3: results

g-estimation of a structural nested model

The g-estimate of ψ and $exp(\psi)$

```
------------------------------------------------------------------------------
| G-estimation: structural nested accelerated failure time models            |
| psi (log time ratio) and time ratio estimates and confidence intervals     |
------------------------------------------------------------------------------
|         Psi (95% CI)                |           Time ratio (95% CI)        |
------------------------------------------------------------------------------
         0.59 (0.36,0.81)                          1.80 (1.43,2.25)
```

Interpretation: every unit of cumulative exposure accelerates you towards your death at 1.8 times your baseline rate[35]

[35] In the special case of a constant hazard, this would imply a hazard ratio of 1.8 per 1 unit increase in cumulative exposure

Longitudinal g-formula

Complications not discussed here

Censoring	All individuals died in our study - censoring requires modifications (see any of the occupational examples)
Late entry	Best to avoid
How to choose models	Not well discussed in the literature

Comparisons

The choice between g-methods

- IPW is usually the easiest approach
- IPW usually fails in occupational studies or with with continuous exposures
- The parametric g-formula is powerful and can nearly always give answers but is full of modeling assumptions (longitudinal)
- G-estimation can work with non-positivity and continuous exposures, but can be strange to interpret directly
- Using more than one method (or doubly robust methods) can help evaluate how much modeling assumptions play a role
- Rule: if you can, use IPW, and if you must, use g-estimation of SNMs. If IPW is too imprecise, then try g-formula.

Reading list

G-formula

ENVR/OCC: A. P. Keil D. B. Richardson (2017) *Environ Health Perspect*;A. P. Keil et al.(2018) *Epidemiology*;A. M. Neophytou et al.(2016) *Epidemiology*

Special, time-fixed case: J. M. Snowden S. Rose K. M. Mortimer (2011) *Am J Epidemiol*;C. J. Muller R. F. MacLehose (2014) *Int J Epidemiol*

Basic implementation: A. P. Keil et al.(2014) *Epidemiology*

Slightly more advanced: S. L. Taubman et al.(2009) *Int J Epidemiol*;D. Westreich et al.(2012) *Stat Med*;R. M. Daniel et al.(2013) *Stat Med*;J. M. Robins M. A. Hernán (2009) *

New-expert level: J. M. Robins M. A. Hernán U. Siebert (2004) *Comparative Quantification of Health Risks: The Global and Regional Burden of Disease Attributable to Major Risk Factors, Genève, Organisation mondiale de la Santé**;J. M. Robins (2008) *Int J Obes (Lond)*

Bible: J. M. Robins (1986) *Math Mod*

Reading list

Structural nested models and g-estimation

ENVR/OCC: A. P. Keil D. B. Richardson M. A. Troester (2015) *Am J Epidemiol*;J. Chevrier S. Picciotto E. A. Eisen (2012) *Epidemiology*;S. Picciotto et al.(2014) *Epidemiology*

Basic implementation: M. A. Hernán et al.(2005) *Pharmacoepidemiol Drug Saf*;J. M. Robins M. A. Hernán (2009)

Slightly more advanced: J. M. Robins (2008) *Int J Obes (Lond)*;J. C. Witteman et al.(1998) *Am J Epidemiol*;N. Keiding et al.(1999) *Biometrics*

New-expert level: J. M. Robins (1992) *Biometrika*

Bible: J. M. Robins (1989) *Health service research methodology: a focus on AIDS*

Resources

These didn't necessarily work for my examples, so I didn't use them

- Stata: gformula, gest (user made programs)
- R: g-formula `https://github.com/ainaimi/pgf`
- SAS: HSPH website for g-formula, g-estimation macros `https://www.hsph.harvard.edu/causal/software/`

The most important points:

- Estimate anything
- Worry about model specification
- Understand causality over time
- Reading list = gateway
- Do, then understand

questions?

Alexander Keil, akeil@unc.edu
Dept. of Epidemiology
University of North Carolina at Chapel Hill


Chevrier, Jonathan, Sally Picciotto, and Ellen A Eisen. “A comparison of standard methods with g-estimation of accelerated failure-time models to address the healthy-worker survivor effect: application in a cohort of autoworkers exposed to metalworking fluids”. In: *Epidemiology* 23.2 (Mar. 2012), pp. 212–9. DOI: 10.1097/EDE.0b013e318245fc06.


Daniel, R M et al. “Methods for dealing with time-dependent confounding”. In: *Stat Med* 32.9 (Apr. 2013), pp. 1584–618.


Greenland, Sander. “For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates”. In: *Eur J Epidemiol* 32.1 (Jan. 2017), pp. 3–20. DOI: 10.1007/s10654-017-0230-6.


Hernán, Miguel A et al. “Structural accelerated failure time models for survival analysis in studies with time-varying treatments”. In: *Pharmacoepidemiol Drug Saf* 14.7 (July 2005), pp. 477–91. DOI: 10.1002/pds.1064.

- Keiding, Niels et al. “The graft versus leukemia effect after bone marrow transplantation: A case study using structural nested failure time models”. In: *Biometrics* 55.1 (1999), pp. 23–28.
- Keil, Alexander P and David B Richardson. “Reassessing the Link between Airborne Arsenic Exposure among Anaconda Copper Smelter Workers and Multiple Causes of Death Using the Parametric g-Formula”. In: *Environ Health Perspect* 125.608-614 (2017).
- Keil, Alexander P, David B Richardson, and Melissa A Troester. “Healthy worker survivor bias in the Colorado Plateau uranium miners cohort”. In: *Am J Epidemiol* 181.10 (May 2015), pp. 762–70.
- Keil, Alexander P et al. “Estimating the Impact of Changes to Occupational Standards for Silica Exposure on Lung Cancer Mortality”. In: *Epidemiology* 29.5 (Sept. 2018), pp. 658–665. DOI: 10.1097/EDE.0000000000000867.

- Keil, Alexander P et al. “The Parametric g-Formula for Time-to-event Data: Intuition and a Worked Example”. In: *Epidemiology* 25.6 (2014), pp. 889–897.
- Muller, Clemma J and Richard F MacLehose. “Estimating predicted probabilities from logistic regression: different methods correspond to different target populations”. In: *Int J Epidemiol* 43.3 (2014), pp. 962–970.
- Neophytou, Andreas M et al. “Occupational Diesel Exposure, Duration of Employment, and Lung Cancer: An Application of the Parametric G-Formula”. In: *Epidemiology* 27.1 (Jan. 2016), pp. 21–8. DOI: 10.1097/EDE.0000000000000389.
- Picciotto, Sally et al. “Hypothetical interventions to limit metalworking fluid exposures and their effects on COPD mortality: G-estimation within a public health framework”. In: *Epidemiology* 25.3 (May 2014), pp. 436–43.

- Robins, J M. “Causal models for estimating the effects of weight gain on mortality”. In: *Int J Obes (Lond)* 32 Suppl 3 (Aug. 2008), S15–41.
- Robins, James M. “A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect”. In: *Math Mod* 7.9 (1986), pp. 1393–1512. ISSN: 0270-0255.
- –. “Estimation of the time-dependent accelerated failure time model in the presence of confounding factors”. In: *Biometrika* 79.2 (1992), p. 321. ISSN: 0006-3444.
- –. *The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. Health service research methodology: a focus on AIDS*. National Center for Health Services Research, US Public Health Service, 1989, pp. 113–159.

- Robins, James M and Miguel A Hernán. *"Estimation of the causal effects of time-varying exposures" in Longitudinal Data Analysis*. Chapman & Hall/CRC, 2009, pp. 553–599.
- Robins, James M, Miguel A Hernán, and Uwe Siebert. "Effects of multiple interventions". In: *Comparative Quantification of Health Risks: The Global and Regional Burden of Disease Attributable to Major Risk Factors, Genève, Organisation mondiale de la Santé* (2004).
- Snowden, Jonathan M, Sherri Rose, and Kathleen M Mortimer. "Implementation of G-computation on a simulated data set: demonstration of a causal inference technique". In: *Am J Epidemiol* 173.7 (Apr. 2011), pp. 731–8. DOI: 10.1093/aje/kwq472.
- Taubman, Sarah L et al. "Intervening on risk factors for coronary heart disease: an application of the parametric g-formula". In: *Int J Epidemiol* 38.6 (Dec. 2009), pp. 1599–611. DOI: 10.1093/ije/dyp192.

- Westreich, Daniel et al. “The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death”. In: *Stat Med* 31.18 (2012), pp. 2000–2009.
- Witteman, J C et al. “G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Heart Study”. In: *Am J Epidemiol* 148.4 (Aug. 1998), pp. 390–401.
- Young, Jessica G, Miguel A Herńan, and James M Robins. “Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data”. In: *Epidemiol Method* 3.1 (Dec. 2014), pp. 1–19.

L=0

		A = 0	A = 1	
Y	0	45	22	67
	1	30	28	58
		75	50	125

L=1

		A = 0	A = 1	
Y	0	28	10	38
	1	22	15	37
		50	25	75

Recall $Pr(Y^g) = \sum_{\ell \in \mathcal{L}} Pr\left(Y \middle| A = g, L = \ell\right) Pr\left(L = \ell\right)$

$\mathcal{L} = \{0,1\}$, Pr(L=0) = 0.625, Pr(L=1) = 0.375

$Pr\left(Y = 1 \middle| A = 1, L = 0\right) = 0.56$

$Pr\left(Y = 1 \middle| A = 1, L = 1\right) = 0.60$

$Pr\left(Y = 1 \middle| A = 0, L = 0\right) = 0.40$

$Pr\left(Y = 1 \middle| A = 0, L = 1\right) = 0.44$

back

L=0

		A = 0	A = 1	
Y	0	45	22	67
	1	30	28	58
		75	50	125

L=1

		A = 0	A = 1	
Y	0	28	10	38
	1	22	15	37
		50	25	75

$$E(Y^1) = Pr\left(Y = 1 \middle| X = 1, L = 0\right) Pr(L = 0) + Pr\left(Y = 1 \middle| X = 1, L = 1\right) Pr(L = 1)$$
$$= 0.56 \times 0.625 + 0.60 \times 0.375 = \mathbf{0.575}$$

$$E(Y^0) = Pr\left(Y = 1 \middle| X = 0, L = 0\right) Pr(L = 0) + Pr\left(Y = 1 \middle| X = 0, L = 1\right) Pr(L = 1)$$
$$= 0.40 \times 0.625 + 0.44 \times 0.375 = \mathbf{0.415}$$

$$(Y^1) - E(Y^0) = \mathbf{0.16}$$

back

Example 4: methods

g-estimation of a structural nested policy model

Say we wish to estimate the effect of a policy, but are uncomfortable with the strong parametric assumptions of the parametric g-formula.

Example 4: methods
g-estimation of a structural nested policy model

We can re-cast “exposure” in the SNM to represent a policy.

```
DATA gestimation_policy;
  SET miners;
  DO psi = 0 TO 1.0 BY 0.01;
  * the 'policy exposure' - you are compliant with the policy (referent)
   if you are below the occupational limit of 0.1, OR if you are not at work,
  Thus, the occupational policy we are testing is "If at work, remain unexposed";
  radlt0_1 = (rad>0.1);
   *blipping down;
    * note that here we define the causal effect of interest;
  * log-linear SNM;
    pyr0 = EXP(psi*radlt0_1)*(outtime-intime);
    OUTPUT;
  END;
```

"If at work, remain below annual exposure occ. limit of 0.1"

Example 4:results

g-estimation of a structural nested policy model

We can re-cast “exposure” in the SNM to represent a policy.

G-estimation: policy structural nested accelerated failure time models	
psi (log time ratio) and time ratio estimates and confidence intervals	
Psi (95% CI)	Time ratio (95% CI)
0.31 (0.09,0.51)	1.36 (1.09,1.67)

"If at work, remain below annual exposure occ. limit of 0.1"
Interpretation: the policy increases the expected lifespan by 36% versus not following the policy

Example 4:results

g-estimation of a structural nested policy model

We can compare this to a similar parameter estimated with the g-formula (using the expected life-span, comparing a pseudo-cohort that follows the policy with one that follows the natural course[36]).

```
* time ratio for comparison with SNM;
PROC SQL;
  TITLE 'Policy time ratio for comparison with structural nested model (SNM TR=1.36)';
  CREATE TABLE a AS
     SELECT
     SUM(1-cum_incidence) AS elifnc FROM pseudo_cohort(WHERE=(intervention=-1));
  CREATE TABLE b AS
     SELECT
     SUM(1-cum_incidence) AS elifexp1 FROM pseudo_cohort(WHERE=(intervention=1));
  SELECT elifexp1/elifnc AS TR FROM a, b;
QUIT;
```

[36] a better comparison might be made with a pseudo-cohort that is always over the limit if at work

Example 4:results

g-estimation of a structural nested policy model

Policy time ratio for comparison with structural nested model

TR
1.4744

Close, but not quite the same! This is an example of when the different parametric approaches to the g-formula and SNMs can lead to different results.[37]

[37] Because this is simulated data, I can tell you the g-formula result is closer to the simulated value. In the course code, we fit a version of the SNM with relaxed parametric assumptions in the exposure model of g-estimation and get a TR that is nearly the same as the simulated value. Thus, our exposure model used in g-estimation may be too parsimonious.