Estimating the effects of time-varying exposures in epidemiologic studies: the Good (methods), the Bad (questions), and the Ugly (Data)

An application of the g-formula to lifestyle and heart disease
Miguel A. Hernán
Harvard School of Public Health
The Good Methods
G-methods

- Methods to appropriately adjust for measured time-varying confounders that are affected by prior exposure
  - G-formula (Robins 1986)
  - G-estimation of structural nested models (Robins 1989)
  - inverse probability weighting of marginal structural models (Robins 1998)

- Confounder-exposure feedback cycles ubiquitous in epidemiologic research
G-methods are not magical
Strong assumptions required

- Exchangeability
  - no unmeasured confounding
- Positivity
- Well-defined interventions

✓ See Robins and Hernán (2008)

- Plus, of course, no measurement error and no model misspecification
The first g-method: The g-formula

- aka the g-computation algorithm formula (Robins 1986)
- Later rediscovered by computer scientists (Pearl et al)

- Under the above assumptions, the g-formula can be used to estimate any average causal effect
- Not a causal method but the causal method
The g-formula

- Dr. Cole has already described it
- Nonparametric method!!
  - Need to estimate the conditional distribution of outcome, confounders...
  - For longitudinal data and/or continuous covariates it requires essentially infinite data and computing time
- Cannot be used in practice
  - Not a trivial issue
Practical solution: The parametric g-formula

1. Estimate the entire likelihood
   - Using parametric models
     - marginal distribution of baseline covariates can be empirically estimated
   - For example, logistic regression for binary variables, linear regression of continuous variables

2. Approximate the integral
   - via Monte Carlo simulation
Should we consider parametric estimation of the g-formula?

- A big chunk of Robins et al’s careers devoted to semiparametric methods
  - IP weighting of MSMs, g-estimation of SNMs
- Inevitable model misspecification viewed as particularly bad for longitudinal data
  - Can lead to catastrophic error propagation
- Now we are talking about going fully parametric? Really?
  - Well, you never know until you try
We tried

- Lifestyle and heart disease
  - Taubman et al. *Int J Epidemiol* 2009

- Antiretroviral therapy and mortality

- When to start antiretroviral therapy
  - Young et al. *Stat Biosci* 2011

- Lifestyle and diabetes
  - Danaei et al. *Epidemiology* 2012 (in press)

- Fish intake and heart disease
  - Lajous et al. *Am J Epidemiol* (under review)
What we learned (I)

- Parametric g-formula estimates more efficient than semiparametric estimates, duh!
  - e.g., IP weighted estimates
- Parametric g-formula is more flexible than any of the other g-methods
- Parametric g-formula is computationally tractable
What we learned (II)

- Most shockingly, and contrary to all expectations, the parametric g-formula yields reasonable and stable estimates
  - As long as enough flexibility is built into the models
  - Not the anticipated crazy answers
- But g-formula estimates are only as good as the questions and the data
The Bad Questions

- Epidemiologists ask questions about the effects of
  - Smoking
  - Diet
  - Alcohol
  - Exercise
  - Weight loss

on the risk of coronary heart disease

- Are these bad questions?
A young couple moves into an apartment and decides to repaper the dining room. They ask the neighbor who has a dining room the same size,

“How many rolls of wallpaper did you buy when you papered your dining room?”

“Seven”, he says

So the couple buys seven rolls of expensive paper, and they start papering. When they get to the end of the fourth roll, the dining room is finished. Annoyed, they go back to the neighbor and say

“We followed your advice, but we ended up with three extra rolls!”

“So”, he says, “that happened to you too.”

(Cathcart and Klein, *Plato and a Platypus Walk into a Bar...*)
Well defined causal questions

- Questions that can be expressed in terms of possibly hypothetical, but well defined interventions

- Think of the hypothetical randomized experiment that would answer the question
  - If you can’t think of one, we have a problem
The Ugly Data
Any data

- Let’s face it
  - Data from epidemiologic studies are ugly

- Data are incomplete
  - Not enough variables (confounders)
  - Not enough people (random error)
  - Not the right people (selection bias)

- Data are measured with error
  - and not frequently enough
The Nurses’ Health Study

- Observational cohort study
  - ~80,000 women with diet, lifestyle data in 1980
- Lifestyle and health information updated by questionnaire every two years
  - Diagnosis of CHD (confirmed by physician)
  - Smoking, weight, diet, alcohol, exercise...
  - Risk factors for CHD
- Arguably one of the best datasets to study the effect of lifestyle on the risk of coronary heart disease (CHD)
“We defined subjects as low risk as those who
1. were not currently smoking
2. had a BMI under 25
3. consumed an average of at least half a drink of an alcoholic beverage per day
4. engaged in moderate-to-vigorous physical activity for at least half an hour per day
5. scored in the highest 40 percent of the cohort for consumption of a diet high in cereal fiber, marine n-3 fatty acids, and folate, with a high ratio of polyunsaturated to saturated fat, and low in trans fat and glycemic load”
Our strategy

1. Define causal questions
   ▪ Specify the hypothetical interventions as explicitly as possible

2. Use the parametric g-formula
   ▪ Because of time-varying confounding
   □ For a detailed description, see
     ▪ Taubman et al. *Int J Epidemiol* 2009
     ▪ Software available from
       www.hsph.harvard.edu/causal
Our attempt to more precisely define the interventions

- Estimated the 20-year CHD risk were the entire population to follow the prescribed intervention (see next slide) beginning at start of follow-up in 1982

- Then compare the estimated CHD risks under each intervention with that under no intervention
Consider 9 hypothetical interventions

1. Avoid smoking
2. Exercise at least 30 minutes a day
3. Keep diet score (described above) in a range corresponding to the top 2 quintiles of the observed data
4. Consume at least 5 grams of alcohol per day
5. Maintain body mass index (BMI) less than 25
6. Interventions 1 - 3 combined
7. Interventions 1 - 3 and 5 combined
8. Interventions 1 - 4 combined
9. Interventions 1 - 5 combined
Our heroic assumptions

☐ No residual confounding
  ■ Given age, parental history of myocardial infarction before 60y, education, husband’s education, ethnicity, hormone use, age at menopause, calendar time, high cholesterol, high blood pressure, diabetes, angina, stroke, coronary revascularization, osteoporosis, body mass index, cigarette smoking, aspirin use, alcohol intake, physical activity, diet score, multivitamin use, fruit/vegetable intake...

☐ No measurement error

☐ No model misspecification
# G-formula estimates

Taubman et al. *Int J Epidemiol* 2009

<table>
<thead>
<tr>
<th>Intervention</th>
<th>20-year Risk</th>
<th>Risk Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0) No intervention</td>
<td>3.68 (3.56, 4.09)</td>
<td>1</td>
</tr>
<tr>
<td>(1) Quit smoking</td>
<td>3.01 (2.86, 3.38)</td>
<td>0.82 (0.78, 0.85)</td>
</tr>
<tr>
<td>(2) Exercise at least 30 minutes per day</td>
<td>2.90 (2.47, 3.60)</td>
<td>0.79 (0.64, 0.92)</td>
</tr>
<tr>
<td>(3) Keep diet score in the top 2 quintiles</td>
<td>3.27 (3.08, 3.68)</td>
<td>0.89 (0.82, 0.95)</td>
</tr>
<tr>
<td>(4) Consume at least 5g alcohol per day</td>
<td>3.19 (2.84, 3.72)</td>
<td>0.87 (0.75, 0.98)</td>
</tr>
<tr>
<td>(5) Maintain BMI less than 25</td>
<td>3.62 (3.45, 4.11)</td>
<td>0.98 (0.93, 1.04)</td>
</tr>
<tr>
<td>(6) “Low-risk” lifestyle (1-3 combined)</td>
<td>2.22 (1.85, 2.74)</td>
<td>0.60 (0.48, 0.70)</td>
</tr>
<tr>
<td>(7) “Low-risk” lifestyle (1-3 and 5 combined)</td>
<td>2.17 (1.78, 2.69)</td>
<td>0.59 (0.47, 0.70)</td>
</tr>
<tr>
<td>(8) “Low-risk” lifestyle (1-3 and 4 combined)</td>
<td>1.88 (1.51, 2.38)</td>
<td>0.51 (0.40, 0.63)</td>
</tr>
<tr>
<td>(9) “Low-risk” lifestyle (1-5 combined)</td>
<td>1.89 (1.46, 2.41)</td>
<td>0.51 (0.39, 0.64)</td>
</tr>
</tbody>
</table>

9-Sept-12 The Good, the Bad, and the Ugly
Interpretation

- 49% of CHD cases attributable to these lifestyle interventions

- Compare with
  - 67% after applying Stampfer et al’s analytic approach to updated NHS data

- Strong effect of lifestyle, though weaker than previously reported
How seriously should we take our estimates?

☐ Good method

☐ But the method’s assumptions are surely violated to some degree

- Bias (of unknown direction and magnitude) because of unmeasured confounding, measurement error, and model misspecification
Unmeasured confounding

- The g-formula appropriately adjusts for *measured* confounding but...
- Surely there is residual confounding by *unmeasured* factors
  - e.g., access to preventive medicine, subclinical disease
- May result in upwards/downwards bias
  - e.g., unmeasured (subclinical) disease would make BMI reduction looks worse, and physical activity increase look better
Measurement error

☐ Surely exposures measured with error

☐ On one hand:
  ▪ If random error, some would argue that most likely direction of bias is towards the null

☐ On the other hand:
  ▪ Because past exposures are also confounders, measurement error results in more residual confounding
  ▪ Bias in either direction
Model misspecification

- Surely our models are misspecified
  - Alternative specifications result in 10% change in estimates

- Correct specification is almost an impossible task:
  - Exposures and confounder measured simultaneously in the same questionnaire
  - Time sequence cannot be discerned
  - Common problem to all “interval” cohorts
And yet...

- Our main contribution is not the use of $g$-methods
  - the parametric $g$-formula

- But the explicit specification of the causal questions
  - which is quite informative in itself
% of subjects whose data is not consistent with the intervention

<table>
<thead>
<tr>
<th>Intervention</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0) No intervention</td>
<td>0</td>
</tr>
<tr>
<td>(1) Quit smoking</td>
<td>30</td>
</tr>
<tr>
<td>(2) Exercise at least 30 minutes per day</td>
<td>99</td>
</tr>
<tr>
<td>(3) Keep diet score in the top 2 quintiles</td>
<td>99</td>
</tr>
<tr>
<td>(4) Consume at least 5g alcohol per day</td>
<td>89</td>
</tr>
<tr>
<td>(5) Maintain BMI less than 25</td>
<td>73</td>
</tr>
<tr>
<td>(6) “Low-risk” lifestyle (1-3 combined)</td>
<td>100</td>
</tr>
<tr>
<td>(7) “Low-risk” lifestyle (1-3 and 5 combined)</td>
<td>100</td>
</tr>
<tr>
<td>(8) “Low-risk” lifestyle (1-3 and 4 combined)</td>
<td>100</td>
</tr>
<tr>
<td>(9) “Low-risk” lifestyle (1-5 combined)</td>
<td>100</td>
</tr>
</tbody>
</table>
By making hypothetical interventions explicit

- We can assess how much our estimates rely on extrapolation from the model
  
  - for some interventions nobody had data consistent with the intervention over the entire follow-up
What else do we learn from making hypothetical interventions explicit?

- Some interventions are still ill-defined

- For example, “Maintain BMI less than 25 starting in 1982”

- Meaning that if your BMI was 30 in 1982, you instantaneously reduce it to 25? How?

  Hernán, Taubman. *Int J Obesity* 2008
The problem of “multiple versions of treatment”
Hernán, VanderWeele. *Epidemiology* 2011

- One can argue that we are really interested in a realistic intervention
  - e.g., encourage exercise and good diet

- Because if we don’t quite know what causal question we are asking
  - Our estimates are hard to interpret
  - Discussion of the merits of any analytic approach is premature
In summary

- Good methods cannot compensate for
  - ugly data
  - bad questions
  - The merits of any method cannot be discussed until the question is specified

- Observational studies may be inadequate for some questions, but how can we even start that discussion if the question is not well defined?
A young couple moves into an apartment and decides to repaper the dining room. They ask the neighbor who has a dining room the same size,

“*How many rolls of wallpaper did you buy when you papered your dining room?*”

“Seven”, he says

So the couple buys seven rolls of expensive paper, and they start papering. When they get to the end of the fourth roll, the dining room is finished. Annoyed, they go back to the neighbor and say

“We followed your advice, but we ended up with three extra rolls!”

“So”, he says, “that happened to you too.”

(Cathcart and Klein. *Plato and a Platypus Walk into a Bar...*)
Key difference between randomized and observational studies

- **Randomized experiments**
  - Question pre-specified in study protocol

- **Observational studies**
  - Question often decided after data have been collected and explored, massaged, tortured...

- This difference may be as important as randomization itself
A proposal: observational studies analyzed like randomized experiments

- Specify the causal question of interest
- Design the protocol
  - eligibility criteria, regimes to be compared, period of follow-up, analytic approach, …
- of a hypothetical randomized experiment to answer the causal question of interest
- Try to emulate such experiment with the observational data + assumptions
Rubin proposed outcome-free design \textit{before} any analysis with observed outcomes

Though emphasis on non-time-varying exposures

“Enthusiasm for the newer statistical tools, while deserved, should be tempered with recognition of the importance of insight, imagination and intimate knowledge of one's field.”