Epidemiologic methods are useless

They can only give you answers

Miguel A. Hernán
Harvard School of Public Health

The problem I'd like to discuss

- Epidemiologists increasingly rely on fancy methods for observational data
  - i.e., any method more complex that a set of 2x2 tables: fancy regressions, marginal structural models, etc.
- Overreliance on methods has led to de-emphasizing the actual research question
- That's bad

Outline of talk: 3 examples

1. When fancy methods are needed to answer the question
   - Antiretroviral therapy for HIV and death
2. When fancy methods are not needed to answer the question
   - Postmenopausal hormone therapy and heart disease
3. When we don't know the question
   - Lifestyle and heart disease

EXAMPLE #1

- Question: What is the effect of combined antiretroviral therapy (cART) on mortality in HIV-infected patients?
- Data: HIV-Causal Collaboration
  - ~60,000 HIV-infected individuals from Europe and the US (PI: Hernan)

Methodological challenge

- Time-varying treatment
  - Individuals start cART at different times
- Time-varying confunders
  - cART initiation depends on the evolving CD4 count, viral load, etc.
- Time-varying confounders are affected by prior treatment
- Conventional methods don't work
  - Robins 1986, Hernán et al. 2004

Let's say we use a Cox model with time-varying treatment

- Treatment as only covariate
  - Hazard ratio: 1.39 (1.25, 1.53)
  - Lots of confounding!
- Add baseline confounders
  - HR: 0.79 (0.70, 0.88)
- Add time-varying confounders too
  - HR: 0.83 (0.74, 0.94)
  - ???

Bias of conventional methods

- Need a "fancy" method: inverse probability (IP) weighting
  - Marginal structural Cox model
- Adjustment for time-varying confounders by IP weighting
  - HR: 0.48 (0.41, 0.57)
  - Consistent with RCT and ecological data

Outline of talk: 3 examples

1. When fancy methods are needed to answer the question
   - Antiretroviral therapy for HIV and death
2. When fancy methods are not needed to answer the question
   - Postmenopausal hormone therapy and heart disease
3. When we don't know the question
   - Lifestyle and heart disease

EXAMPLE #2

- Question: What is the effect of postmenopausal hormone therapy on risk of coronary heart disease in postmenopausal women?

Answers (shocking discrepancy)

- Observational studies
  - >30% **lower risk** in current users compared with never users
    - e.g., HR 0.68 in Nurses' Health Study (Grodstein et al. *J Women's Health* 2006)
- Randomized trial
  - >20% **higher risk** in initiators compared with noninitiators
    - HR 1.24 in Women's Health Initiative (Manson et al. *NEJM* 2003)

WHI: ITT effect estimates

Hazard ratio (95% CI) of CHD

- Overall 1.23 (0.99, 1.53)
- Years of follow-up
  - 0-2 1.51 (1.06, 2.14)
  - >2-5 1.31 (0.93, 1.83)
  - >5 0.67 (0.41, 1.09)
- Years since menopause
  - <10 0.89 (0.54, 1.44)
  - 10-20 1.24 (0.86, 1.80)
  - >20 1.65 (1.14, 2.40)

Why did observational studies get it "wrong"?

- Popular theory: residual confounding
  - insufficient adjustment for lifestyle and socioeconomic indicators
  - Corollary: causal inference from observational data is a hopeless undertaking
- An alternative theory: Observational and randomized studies asked different **questions**

Randomized experiment

- First state your question, then decide your analytic approach
  - **Explicit** causal question: what is the CHD risk in women who **initiate** hormone therapy?
- Design and analysis:
  - Women randomly assigned to initiation of hormone therapy or placebo
  - Analytic approach: Compare risk between women who initiate and do not initiate therapy (Intention-to-treat)

Observational studies

- First decide your analytic approach,
  - Compare risk between women who currently use therapy and those who never used it
- Then find out the question you are answering?
  - **Implicit** causal question: e.g., what is the risk among women who continue to use hormone therapy?

"Current vs. never" contrast does not address a relevant question

- Consider a woman wondering whether to start hormone therapy
  - The current vs. never does not provide the information she needs
- Consider a woman wondering whether to stop hormone therapy
  - The current vs. never does not provide the information she needs

Our strategy

- Use the observational data to answer same question as randomized experiment
  - Re-analyze observational studies to estimate the observational analog of the ITT effect of therapy initiation
- Then compare both set of estimates
- For a detailed description see
  - Hernán et al. *Biometrics* 2005
  - Hernán et al. *Epidemiology* 2008

Data: 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
  - Use of hormone therapy
  - Diagnosis of CHD (confirmed by physician)
  - Risk factors for CHD
- We use this observational study to emulate a "trial" of hormone therapy
  - Starting in period before 1984 questionnaire

Protocol of the NHS "trial" Interventions and Eligibility criteria

- Treatment regimes

1) Initiation of oral estrogens plus progesterone therapy at baseline
2) No hormone initiation at baseline

- Similar eligibility criteria as randomized experiment
  - Including washout interval: no hormone use in 2-year period before baseline

Protocol of the NHS "trial" Baseline and Follow-up

- Baseline:
  - Initiators: month of initiation in 2-yr period before the 1984 questionnaire
  - Non initiators: average baseline month among initiators in same period
- Follow-up
  - From baseline to CHD diagnosis, death from other causes, loss to follow-up, or June 2000, whichever came first

The NHS "trial" Summary

- The NHS "trial" can be viewed as a nonrandomized, nonblinded trial that mimics the eligibility criteria, definition of start of follow-up, and treatment arms of the WHI randomized trial
- Some differences
  - distribution of baseline characteristics
    - shorter time since menopause in NHS
  - Longer follow-up in NHS than in WHI

Protocol of the NHS "trial" Intention to treat (ITT) principle

- Compare the CHD risk between initiators and noninitiators of hormone therapy at baseline
- Regardless of future use during the follow-up
- Observational analog of the ITT effect
  - Using Cox model like the WHI did

The NHS "trial" Non randomized after all

- All confounders have to be appropriately measured and adjusted for in the model
- We included the following baseline variables
  - Age, past hormone use, parental history of myocardial infarction before 60y, education, husband's education, ethnicity, 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, and fruit/vegetable intake

No "fancy" methods

- Just a Cox model
  - Not even time-varying variables
- But question has been changed
  - from comparison of current users vs. never users
  - to comparison of initiators vs. noninitiators

  of hormone therapy

One more thing: A sequence of NHS "trial**s**"

- We started the NHS "trial" during the period before the 1984 questionnaire but there is nothing special about the 1984 questionnaire
- We also started a "trial" in the periods before the 1986, 1988, … 1998 questionnaires
  - A sequence of nested "trials"
- Each woman may participate in a maximum of 8 trials

The NHS "trial**s**"

- For each trial
  - Follow-up started at the trial-specific baseline and ended at diagnosis of CHD, death, lost to follow-up, or June 2000
  - Eligibility criteria applied at baseline
- We pool data across "trials" to obtain an effect estimate with a narrower confidence interval
  - Robust variance because of within-subject correlation

Results
Women eligible for NHS "trials"

- 34,472 women contributed to trials
  - 1,021 CHD cases
- Pooling over "trials"
  - On average, each woman participated in 4.4 trials
  - 152,479 participants
  - 6,602 initiators
  - 3,597 CHD cases

ITT effect estimates

	WHI	NHS
Overall	1.23 (0.99, 1.53)	1.05 (0.82, 1.34)
Years of follow-up		
0-2	1.51 (1.06, 2.14)	1.43 (0.92, 2.23)
>2	1.07 (0.81, 1.41)	0.91 (0.72, 1.16)
Years since menopause		
<10	0.89 (0.54, 1.44)	0.88 (0.63, 1.21)
10-20	1.24 (0.86, 1.80)	1.13 (0.85, 1.49)
>20	1.65 (1.14, 2.40)	--

When same **question** is asked

- No shocking observational-randomized discrepancies
  - though wide CIs in both studies
- **No fancy methods required**
- What about the popular hypothesis? Any residual confounding?
  - Probably, but insufficient to explain the original discrepancy

Clarification:

- In the interest of time, I simplified Case study #2 to make a simple point
  - *Question matter more than methods*
- Full analysis is more nuanced
- For example, we used an ITT approach because we wanted a direct comparison with the WHI estimates

But ITT analyses are problematic

- ITT effect affected by adherence
  - Imperfect adherence in both randomized and observational studies
- ITT inappropriate for safety outcomes
- We also conducted IP weighted analyses to adjust for noncompliance
  - Still no randomized-observational discrepancies
  - Toh et al. *Ann Intern Med* 2010

Outline of talk: 3 examples

1. When fancy methods are needed to answer the question
   - Antiretroviral therapy for HIV and death
2. When fancy methods are not needed to answer the question
   - Postmenopausal hormone therapy and heart disease
3. When we don't know the question
   - Lifestyle and heart disease

EXAMPLEs #1 and #2

1. Question of interest was well defined
   - it was a matter of appropriately analyzing the observational data
2. Magnitude of the effect RCT was approximately known from RCTs
   - We had a benchmark
- What if neither of these 2 elements are present?

EXAMPLE #3
Lifestyle and heart disease

- Less well-defined causal questions
  - What is the causal effect of smoking, alcohol, physical activity, diet, and body mass index (BMI) on the risk of CHD?
  - Need to express this question in terms of hypothetical interventions or counterfactual contrasts
- Unknown answer
  - No large, long-term randomized trials of lifestyle with full adherence to the relevant interventions

The Nurses' Health Study
Stampfer et al, *NEJM* 2000

- "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 question and assumptions
   - Specify the hypothetical interventions as explicitly as possible
2. Use a "fancy" method
   - 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 assumptions

- No residual confounding
  - Given same variables listed before
- No measurement error
- No model misspecification

- We'll discuss each of them later

Estimates from the NHS

Taubman et al. *Int J Epidemiol* 2009

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

Interpretation

- 49% of CHD cases attributable to these lifestyle interventions
- Compare with
  - 82% in Stampfer et al (2000)
  - 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?

- Analytic approach is now consistent with the assumptions
- But 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:
  - possible attenuation of the effect
  - e.g., if measurement error for diet and average 2-year change in diet are of similar magnitude
- On the other hand:
  - Because past exposures are confounders, measurement error results in more residual confounding

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 "fancy" methods
  - e.g., the parametric g-formula
- But the explicit specification of causal question of interest
  - By making the hypothetical interventions explicit

% of subjects whose data is not consistent with the intervention

Intervention	%
(0) No intervention	0
(1) Quit smoking	30
(2) Exercise at least 30 minutes per day	99
(3) Keep diet score in the top 2 quintiles	99
(4) Consume at least 5g alcohol per day	89
(5) Maintain BMI less than 25	73
(6) "Low-risk" lifestyle (1-3 combined)	100
(7) "Low-risk" lifestyle (1-3 and 5 combined)	100
(8) "Low-risk" lifestyle (1-3 and 4 combined)	100
(9) "Low-risk" lifestyle (1-5 combined)	100

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 and Taubman. *Int J Obesity* 2008

The problem of "multiple versions of treatment"

- 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

Two stages in causal inference from observational data

1. Formulate a reasonably well-defined causal question
2. Propose an answer by combining
   - Available data
   - Untestable assumptions
   - Appropriate analytic method

- Often discussions (and lectures in courses) revolve exclusively around Stage #2

Key difference between randomized and observational studies

- Randomized experiments
  - Question and analytic approach pre-specified in study protocol
- Observational studies
  - Question and analytic approach often decided after data have been collected and explored, massaged, tortured…
- This difference may be as important as randomization itself

A way forward: observ. studies analyzed like random. 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**

Conclusions

- Asking the same question decreases the randomized-observational discrepancies
- Observational studies may still be inadequate for some questions, but how can we even start that discussion if the question is not well defined?
  - The merits of any analytic approach cannot be discussed until the question of interest is clearly specified

Conclusions

- The complexity of causal questions is often overlooked
  - especially when dealing with complex longitudinal data
- Emphasis on sophisticated epidemiologic methods becomes a red herring
  - Fancy methods will give us answers, but for what question?