Defining and estimating intervention effects
for groups who will develop an auxiliary outcome

Marshall M. Joffe

Department of Biostatistics and Epidemiology
University of Pennsylvania
Motivating example: nephrology

Nephrologists frustrated by the lack of good means to modify the course of end-stage renal disease (ESRD)

Suggested methods to modify this course disappointing

- better/ more aggressive control of blood pressure
- higher dose of dialysis

Frustration has led some to hypothesize:

Time period before ESRD may provide a window of opportunity to improve outcomes among ESRD patients

Hope: interventions or treatments applied before ESRD may affect the course of ESRD
Background: study of chronic renal insufficiency (CRI)

CRI: reduced renal function, no ESRD yet (requiring dialysis or transplant)

Chronic Renal Insufficiency Cohort Study (CRIC)

Multicenter observational study

Goal: recruit about 3000 participants in 7 centers

In recruitment phase
Main endpoints:

- ESRD
- cardiovascular (CV) events

Wanted to study these issues as part of “ancillary study” application

This simple hypothesis surprisingly hard to formalize

Consider

- various ways to formalize hypothesis
- whether and when hypothesis so formalized is of interest
“Philosophical” talk: speak about what is of interest, little about how to estimate

Illustrate with simple example:

binary treatment $A$: aggressive management of blood pressure

outcome $Y$: development of myocardial infarction (MI)

auxiliary variable $S$: development of ESRD

implications extend beyond ESRD example
Naive approach

Compare subjects who develop ESRD, those who do not

compute difference in probabilities:

$$E(Y|S=1,A=1) - E(Y|S=1,A=0)$$

problem: not defined as comparison of what would happen to individual or group under different circumstances
potential outcomes model (Neyman, 1923; Rubin, 1974, etc.)

$Y^a$: outcome that one would see if receive treatment level $a$

if receive other treatment level, $Y^a$ counterfactual

causal effects: comparisons of $Y^1$, $Y^0$ for individual or group

e.g., $E(Y^1) - E(Y^0)$: comparison of outcome in same subjects under different conditions
rephrase naive approach in terms of potential outcomes:

\[ E(Y^1|S=1,A=1) - E(Y^0|S=1,A=0) \]

\[ = E(Y^1|S^1=1,A=1) - E(Y^0|S^0=1,A=0) \]

problem: even if treated and untreated groups comparable at beginning of study, group membership may be affected by treatment in randomized trial, above estimand equals

\[ E(Y^1|S^1=1,A=1) - E(Y^0|S^0=1,A=0) \]

\[ = E(Y^1|S^1=1) - E(Y^0|S^0=1) \]

comparison of outcomes under different groups
this is causal in sense that it correctly describes the impact of treatment on conditional distribution of MI/outcome given auxiliary outcome $S$ (ESRD)

not a comparison of impact of treatment for given group

nor for individuals

unsatisfying

there is some value in specifying when the effect of treatment on a distribution can be estimated correctly

in observational study, will not be true that

$$E(Y^1|S^1=1,A=1) - E(Y^0|S^0=1,A=0) = E(Y^1|S^1=1) - E(Y^0|S^0=1)$$

conditions for estimating effects on distributions require further development
numerical example:

aggressive treatment determines who gets ESRD

no effect on MI for any subject

| $S^0$ | $S^1$ | $pr(Y^0=1|S^0,S^1)$ | $pr(Y^1=1|S^0,S^1)$ | N |
|-------|-------|----------------------|----------------------|---|
| 0     | 0     | 0.1                  | 0.1                  | 500 |
| 1     | 0     | 0.2                  | 0.2                  | 300 |
| 1     | 1     | 0.3                  | 0.3                  | 200 |
consider conditional distributions given exposure, outcome

| $a$ | $S^a$ | $N$   | $pr(Y^a=1|S^a)$ |
|-----|-------|-------|-----------------|
| 0   | 0     | 500   | 0.1             |
| 0   | 1     | 500   | 0.24            |
| 1   | 0     | 800   | 0.1375          |
| 1   | 1     | 200   | 0.3             |

Aggressive treatment reduces probability of ESRD from 50% to 20%

Among subjects with ESRD, probability of MI increases from 24% to 30%
effects on distributions above can be estimated from randomized trial data

| $A$ | $S$ | $N$   | $pr(Y=1|S,A)$ |
|-----|-----|-------|----------------|
| 0   | 0   | 250   | 0.1            |
| 0   | 1   | 250   | **0.24**       |
| 1   | 0   | 400   | 0.1375         |
| 1   | 1   | 100   | **0.3**        |

clearly doesn’t represent desired effect

now have fewer subjects with ESRD

sicker
other options

principal stratification (Frangakis and Rubin)

consider subjects who would get ESRD whether or not they get aggressive treatment

| $S^0$ | $S^1$ | $pr(Y^0=1|S^0,S^1)$ | $pr(Y^1=1|S^0,S^1)$ | N  |
|-------|-------|----------------------|----------------------|----|
| 0     | 0     | 0.1                  | 0.1                  | 500|
| 1     | 0     | 0.2                  | 0.2                  | 300|
| 1     | 1     | 0.3                  | 0.3                  | 200|

each row is principal stratum

based on cross-classification of subjects by \( \{S^0,S^1\} \)
meaningful causal estimand

interpretation of question
problems:

can’t identify subjects in stratum in general without additional assumptions

principal strata partially identified

e.g., subject treated nonaggressively with ESRD (i.e., $A=0$, $S=1$) may be in either of 2 strata:

since $S^0 \in \{0,1\}$, $\{S^0, S^1\} \in \{(1,0),(1,1)\}$
one assumption which fully identifies principal stratum of interest in treated subjects: monotonicity

aggressive treatment never causes ESRD: \( S^1 \leq S^0 \)

true in artificial example considered here

consider later why not credible assumption here
effects defined within strata defined by classification of subjects by single auxiliary variable $S^0$ or $S^1$

effects for subjects who would develop ESRD if don’t receive aggressive treatment

| $S^0$ | $pr(Y^0=1|S^0)$ | $pr(Y^1=1|S^0)$ | N  |
|-------|-----------------|-----------------|----|
| 0     | 0.1             | 0.1             | 500|
| 1     | 0.24            | 0.24            | 500|

effects for subjects who would develop ESRD if receive aggressive treatment

| $S^1$ | $pr(Y^0=1|S^1)$ | $pr(Y^1=1|S^1)$ | N  |
|-------|-----------------|-----------------|----|
| 0     | 0.1375          | 0.1375          | 800|
| 1     | 0.3             | 0.3             | 200|
here: same as principal stratification effect

either is interpretation of researcher’s question

compare $E(Y^a | S^{a'} = s)$ for different $a$
again, can’t in general identify strata in entire population

however, subjects in stratum identifiable among subjects exposed to desired treatment level $a'$

in randomized trial, we have:

$$E(Y^a|S^{a'}=s)=E(Y^a|S=s,A=a')$$

so can identify group for whom effect defined

generalizations available to observational studies if make assumption that treatment assignment ignorable
can define effects in terms of identifiable groups, based on observed variables:

effects for subjects who developed ESRD on non-aggressive treatment:

| $S$ | $\Pr(Y^0=1|S,A=0)$ | $\Pr(Y^1=1|S,A=0)$ | N |
|-----|---------------------|---------------------|---|
| 0   | 0.1                 | 0.1                 | 250 |
| 1   | 0.24                | 0.24                | 250 |

effects for subjects who developed ESRD on aggressive treatment:

| $S$ | $\Pr(Y^0=1|S,A=1)$ | $\Pr(Y^1=1|S,A=1)$ | N |
|-----|---------------------|---------------------|---|
| 0   | 0.1375              | 0.1375              | 400 |
| 1   | 0.3                 | 0.3                 | 100 |

estimation can’t be based on comparing aggressively treated to others among those with ESRD
estimation can be based on following idea (Palmgren, Robins):

in randomized trial, the latent potential outcome does not affect treatment group

e.g., \( pr(A=1|Y^0)=pr(A=1) \)

extensions to observational studies

propose model for effect of treatment:

\[
E(Y|A,S)-E(Y^0|A,S)=A(1-S)\Psi_0 + AS\Psi_1
\]

(Can consider other link functions)

\( \Psi_0 \) is effect of aggressive treatment for treated subjects who do not get ESRD

\( \Psi_1 \) is effect of aggressive treatment for treated subjects who do not get ESRD
if model is true, can compute

\[ E(Y^0|A,S) = E(Y|A,S) - A(1-S)\Psi_0 - AS\Psi_1 \]

should be independent of \( A \) marginally

like G-estimation, structural nested models, but allow effects to be defined conditional on variables not measured at time of treatment

exploit this to estimate parameters

should require some baseline covariate \( X \) which predicts \( S \): i.e.,

\[ pr(S|X,A=1) \neq pr(S|A=1) \]

can also be viewed as effect modification by \( S \)
explanatory flavor to analysis/model:

effects are defined for groups not identified at time of treatment decision

thus, model can’t be used directly to predict effect of treatment at time of decision (can be used in more complicated fashion)

these effects, which have already happened to defined subgroups, explain differences between randomized groups
can estimate effects of aggressive treatment for subjects likely to develop ESRD (if given particular treatment):

define group based on covariates:

let $\mu^a(X) = pr(S^a = 1|X)$ (Principal score (Hill et al., 2002))

under ignorable treatment assignment/no unmeasured confounders $(pr(A = a|S^a, X) = pr(A = a|X))$, can estimate as $\hat{\mu}^a(X) = pr(S = 1|X, A = a)$

observable variable
estimate $E\{Y^a|\mu^a(X)\}$, compare for different $a$

estimate as $E\{Y|\hat{\mu}^a(X), A=a\}$ in randomized trial

conventional statistical methods apply

if model $E\{Y|\hat{\mu}^a(X), A=a\}$, look for effect modification by $\hat{\mu}^a$
suppose follows linear model: 

$$E\{Y^1|\mu^{a'}(X)\} - E\{Y^0|\mu^{a'}(X)\} = \alpha_0 + \alpha_1 \mu^{a'}(X)$$

extrapolate to hypothetical subgroup where all subjects will get ESRD even if treated aggressively; \(\mu^{a'}(X) = 1\)

obtain \(\alpha_0 + \alpha_1\)

tempting to interpret as effect among subjects who would get ESRD if given treatment level \(a'\)

need additional conditions for this to be true

extrapolation will require that some covariate predicts auxiliary variable \(S^{a'}\)
can estimate effects of aggressive treatment in subjects likely to develop ESRD whether or not treated aggressively

\[ E\{Y^a|\mu^0(X),\mu^1(X)\} \]

again, identified from observed data using conventional statistical methods

has flavor of principal stratification but is not
last option: conventional statistical methods, ignore ESRD/auxiliary variable $S$

look for effect modification by baseline covariates
what is preferable?

partly philosophical question

for whom do we want to estimate causal effects?

Purpose of modeling:

• practical

• explanatory
practical purposes:

guide interventions in populations similar to those studied

easiest to use methods based only on baseline covariates

harder to use methods based on other observed covariates

harder still to use methods which condition on unobservable quantities:
calculations more involved
explanatory purposes: understanding why/how some outcomes come to pass

can use methods which condition on post-intervention variables

experimentalist vs. historian paradigm

question: how does conditioning on unobserved ESRD status aid in mechanistic/biochemical understanding of mechanisms of treatment effect?
Alternative approach to mechanisms: direct and indirect effects

requires positing more potential outcomes, considering both A and S potentially modifiable

$Y^{a,s}$: outcome one would see if received level $a$ of treatment, level $s$ of ESRD

Direct effect:

comparison of $Y^{a1,s}$, $Y^{a0,s}$

Indirect effect:

harder to define: (see definitions by Pearl, Robins)
Consider several explanations; which is most satisfying?

1. The effect of aggressive treatment of blood pressure on MI is greater for people who subsequently develop ESRD (observed auxiliary stratification);

2. The effect of aggressive treatment of blood pressure on MI is greatest for people who would develop ESRD only if not treated aggressively (principal stratification);

3. Aggressive treatment of blood pressure prevented the development of ESRD in some subjects (one principal stratum). This, in turn, prevented the development of MI for some subjects. Thus, ESRD mediated in part the effect of aggressive management of blood pressure on MI; and

4. Aggressive treatment of blood pressure prevents the development of ESRD in some subjects. This, in turn, prevents the development of MI for some subjects. Thus, ESRD mediates in part the effect of aggressive management of blood pressure on MI.
1st 2 explanations: for whom does treatment work for?

Latter 2: how does treatment work?

Although complementary, prefer latter (where supportable)

helps to make statements like “treatment works because ...”
Direct and indirect effects defined in interventionist terms:

meaningful to discuss even if second factor (e.g., kidney function/ESRD) not under control of investigator

need to consider hypothetical intervention/thought experiment

e.g., replace poor kidney by genetically identical kidney (e.g., from identical twin) in perfect operation

not feasible with current technology

somewhat idealized

explanation testable/parameters estimable in animals

thus, in principle, have possibility of falsification/confirmation of explanation
Estimation in practice complicated

Can be based on

1. ignorability of both primary treatment $A$ and auxiliary factor $S$

or

2. assumed model for joint effect, ignorability only of primary treatment $A$
   (see Robins and Greenland, 1996)
1. ignorability of both primary treatment $A$ and auxiliary factor $S$; e.g.,

$$pr(A|X,Y^{a,s})=pr(A|X)$$

$$pr(S|A,X,L,Y^{a,s})=pr(S|A,X,L)$$

may need to include post-treatment ($A$) factors $L$ in second expression

second condition is hard to consider when $S$ is not under human control

Processes leading to ESRD poorly understood

Hard to render judgments about whether subjects who develop ESRD at some point are otherwise comparable to others
2. assumed model for joint effect, ignorability only of primary treatment $A$
   (see Robins and Greenland, 1996)

   dependant on modeling assumptions; no nonparametric identification
   estimation will generally be less efficient than under full ignorability
   method essentially uses $A$ as instrument for itself and auxiliary $S$
contrast between situation and other situations for which principal stratification proposed

“Censoring by death” (Robins, 1986; Frangakis and Rubin; Zhang)

Causal contrast $Y^1$ vs. $Y^0$ only defined for subjects for whom $S^1=S^0=0$ (i.e., subjects who don’t die)

randomized trials with noncompliance:
arguable that causal effect of treatment only of interest for subjects who accept encouragement $S^1=1, S^0=0$ (Angrist et al., 1996)

Imbens et al. (1997), others look at effect of randomization within principal strata: effect modification by principal strata

under exclusion restriction, interested in effect primarily in stratum of compliers
Here, outcome (CVD) is well-defined and meaningful whether or not one develops ESRD.

Reason for principal stratification less compelling.
additional issues/extensions

both $Y$ and $S$ are censored failure-time outcomes

identification of subgroups for whom effects defined harder

could simplify by taking $Y$ as indicator of failure/CVD at end of fixed interval

$S$ as indicator of ESRD at some time point

lose some important information by dichotomizing

if don’t dichotomize: if look at effect modification by $S$,

$S$ missing if censoring

stratifying on risk scores or baseline covariates simpler
treatment is actually time-varying

increases # of potential outcomes to be considered

subsets to be used in analysis

if formulate as principal stratification problem, dimension of problem can increase much faster than for methods which condition on observed outcomes
censoring by death:

some subjects will die before developing ESRD

raises (at least) 2 issues:

1. defining estimands

now CVD, ESRD outcomes ill-defined for subjects who die

principal stratification may be more compelling approach

\( D^a \): indicator of whether die under treatment \( a \)

compare \( E(Y^a | D^0 = D^1 = 0) \)
can define estimands in conjunction with other approaches:

dual principal stratification:

\[ E(Y^a | D^0 = D^1 = 0, S^0 = S^1 = 1) \] (effect for subjects who would develop ESRD whether or not treated aggressively)

\[ E(Y^a | D^0 = D^1 = 0, S^1 = 1) \] (effect for subjects who would develop ESRD whether or not treated aggressively)
2. if aggressive treatment prevents death, direction of bias from naive analysis not clear

with monotonicity (no death), analysis made aggressive treatment look bad (more people at high risk survive)

now, no monotonicity overall: some subjects who would have died if not treated aggressively would live and develop ESRD

these are likely sicker people who would be more likely to develop CVD

biases comparison of people who would have lived under both treatments
acknowledgments

Dylan Small
Department of Statistics, University of Pennsylvania

Zhen Chen
Michael Elliott
Tom Ten Have
Division of Biostatistics, University of Pennsylvania

Chi Hsu
Department of Medicine, University of California, San Francisco