Appearance versus reality: on reconciling the many faces of causal effects estimated in the medical literature.

Bianca De Stavola (LSHTM, UK)
Els Goetghebeur (Gent, B)
Saskia Le Cessie (Leiden, NL)
Erica Moodie (Mc Gill, Ca)
Ingeborg Waernbaum (Umea,S)

Joint work with Marie Eriksson, Arnout Van Messem, Stijn Vansteelandt & Machteld Varewyck

Victoria, July 12, 2016

Happy 19th Birthday Malala Yousafzai !

Photo by Awais Azad - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=49907427

Causal inference on the rise

- Avaiable state of the art methods - with adapted software - is exploding
- Sophisticated methods are entering 'mainstream' use
- Application is demanding at the conceptual and technical level
- Adaptation in complex settings (EHR,...) with missing data, requiring model selection etc. not well understood
- Ever more ambitious in types of questions to answer
  total causal effect - mediation - optimal dynamic treatments

Back to basics, robust, transportable meaning: Can less be more?

Highlight some challenges and possible ways of handling them
in the point exposure set-up

I. Phrasing the causal question: (internal validity)

Contrast potential outcome distribution for exposure A vs. B

- Clear (enough) on nature of exposure A [1]
- Clear (enough) on the potential outcomes $Y(\mathfrak{a})$
- Clear causal effect estimand: for what population $E(Y(\mathfrak{a}))$

Promotion of Breastfeeding Intervention Trial PROBIT (Kramer et al., 2001):

- (Cluster) randomised pregnant women through educational program on uptake of breastfeeding at birth
- Some 17,044 healthy mothers with full term live singleton births in Belarus (9,565 active arm; 7,479 placebo arm).
- Our focus on point exposure 'started breastfeeding' and outcome 'weight at 3 months'

  simulation study mimicked real data

[1] Vandenbroucke et al., 2016, IJE

Well defined exposure?

'Starting BF' is well defined as exposure (narrow window), but...

- entails a distribution of breast feeding patterns
  in terms of duration, timing, mode, etc.
- We study whatever form (distribution) it takes in our study
- For meaning/understanding + transportability consider
  - form of prescription ['per protocol']
  - form of uptake: when and how by whom ['compliance']

II. Data structure and assumptions justified in context

1. To define the question: *What if* exposure A vs. B
   - Positivity
   - No interference
   - Consistency

e.g. No interference: one individual's treatment effect does not depend on the treatment status of others

TRUE : 'no interference' is likely met because breastfeeding one baby is unlikely to affect the weight of another'

FALSE : 'no interference' is violated: a baby without beast feeding
⇒ more susceptible to infectious disease
⇒ more infection for neighbouring babies hence lower W3

2. To help answer the question from observational data

2a. Fundamentally

- No unmeasured confounders - **L** measured confounders
- Instrumental variable(s) Z

- Choice of **L** in practice (EHR) ?
- with missing data, measurement error and over fitting?
- Internal vs. external validity [2]

2b. Modeling assumptions [checking?]

- Structural model: for potential outcomes (e.g. MSM)
- Association models (testable !)
  - Outcome regression model
  - Propensity of *treatment* regression model

[2] Keiding and Louis, 2016, RSS-A

III. Classes of estimation methods

Assuming 'No unmeasured confounders'

- Direct confounder adjustment
  Outcome regression/stratification/matching based
  (may or may not involve propensity score as an aid)
- Inverse probability of treatment: incl. propensity score
- Double robust methods [3]: combines the above

Using outcome working model

$$E(Y|A=c,\mathbf{L}) = m(c,\mathbf{L};\alpha,\boldsymbol{\beta})$$

and a propensity score working model

$$P(A=c|\mathbf{L}) = h(c,\mathbf{L};\alpha_c^*,\boldsymbol{\beta}_c^*)$$

$$\hat{E}\{Y(c)\} = \frac{1}{n}\sum_{i=1}^{n} m(c,\mathbf{L}_i;\widehat{\alpha},\widehat{\boldsymbol{\beta}}) + \frac{1}{n}\sum_{i=1}^{n}\frac{A_{ic}}{h(c,\mathbf{L}_i;\widehat{\alpha}_c^*,\widehat{\boldsymbol{\beta}}_c^*)}\left\{Y_i - m(c,\mathbf{L}_i;\widehat{\alpha},\widehat{\boldsymbol{\beta}})\right\}$$

[3] Bang and Robins, 2005, Biometrics

Confounders and effect modifiers in L

- A=1 may differ from observed group A=0 in distribution of **L** prognostic factors for $Y(0)$ (baseline characteristics)
- Assume: Conditional on measured **L**, A=1 group and A=0 group have exchangeable $(Y(0), Y(a))$.

regress Y on **L** in $\{A = 1\}$	$-> F_1(y\|\ell)$
regress Y on **L** in $\{A = 0\}$	$-> F_0(y\|\ell)$

$F_1(y|\ell) \leftrightarrow F_0(y|\ell)$ contrast reflects causal effect of $\mathfrak{a}$ for given **L**.

Outcome regression

$Y(\mathfrak{a}) \coprod A|L \quad \forall \mathfrak{a} \quad \Rightarrow$

$$\{Y|L, A=a\} = \{Y(a)|L, A=a\} \overset{d}{=} \{Y(a)|L\}$$

Hence simply regress Y on L in several A-defined strata
to infer the population distribution of $Y(a)$ conditional on L.

regress Y on L in $\{A=1\}$	$->f_1(y|\ell)$
regress Y on L in $\{A=0\}$	$->f_0(y|\ell)$

Challenges:

- With 'high' dimension of ℓ : confidence in a correct model
- $L-$distribution for (non)treated does not overlap ($\pm$)
  e.g. in the young and fit you may find no statin users
- $E(Y|L, A=1) - E(Y|L, A=0) =$
  $E(Y(1)|L) - E(Y(0)|L) = \psi(L)$ i.e. may differ over L

Confounders and population specific summary

Summarize this $F_1(y|\ell) \leftrightarrow F_0(y|\ell)$ contrast for target population :

- the **study population** :
  - ACE: E (Y(1)) - E (Y(0)) and $\hat{E}(Y(a)) = \frac{1}{n}\sum_{i=1}^{n} F_a(y|L_i)$
  - ACE$_1$: $E(Y(1)|L=1) - E(Y(0)|L=1)$; $\{L:\}$ education level
- **the treated** study population:
  - ATT$_1$: $E(Y(1)|A=1) - E(Y(0)|A=1)$ using $\frac{1}{n_1}\sum_{i:A_i=1}\hat{F}_1(y|L_i)$ etc.
- **extrapolated target population** with own **L**− distribution: ACE$_{w(\ell)}$: E$_{w(\ell)}$ (Y(1)) - E$_{w(\ell)}$ (Y(0))
- in **potential principal strata** [4] (following randomisation, IV) CACE: $E(Y(1)|(A(1)=1, A(0)=0)) - E(Y(0)|(A(1)=1, A(0)=0))$

[4] Frangakis and Rubin, 2002, BICS

Simulation study mimics PROBIT

Figure 1: Data generating diagram, in red the causal effect of interest

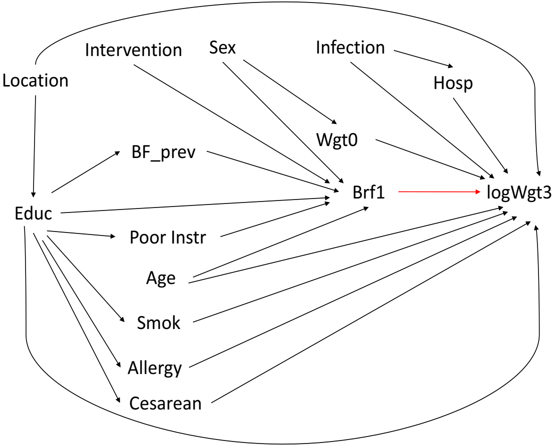


Table 1: Summary of estimated causal effects

Question	(a)		(b)				(c)	
Estimand	ACE		ACE_0		ACE_1		ATT	
True value	148.27		210.06		112.69		124.99	
Estimate	$\widehat{\text{ACE}}$	SE	$\widehat{\text{ACE}_0}$	SE	$\widehat{\text{ACE}_1}$	SE	$\widehat{\text{ATT}}$	SE
Crude regression	253.42	5.45	305.78	8.65	210.40	7.05		
Regression adjustm.	151.03	1.85	212.74	2.91	116.14	2.25	128.31	2.26
Regression with PS	155.48	1.98			123.05	2.53	134.94	5.99
PS stratification	157.49	6.65	218.28	8.41	121.37	9.12	121.53	5.53
PS matching	154.46	3.96	207.62	5.28			131.01	6.34
PS IPW	147.16	2.44	212.11	3.09	111.76	3.02	119.47	4.01
IV (simple)	136.00	29.38	225.52	44.81	81.18	38.28	136.00	29.38
IV (with confounders)	152.44	10.79	199.87	17.20	124.61	13.57	152.14	10.81

Missing data and variable selection in Riksstroke - QOC

- n patients treated in one of the m centers,
- p measured characteristics $\mathbf{L}$

Assuming ‘no unmeasured confounders’:

$$Y(c) \perp\!\!\!\perp C|\mathbf{L},$$

we can estimate the directly standardized risk $E(Y(c))$ as:

$$E(Y(c)) = E\left(E(Y|\mathbf{L}, C = c)\right)$$

Model for Y indicating 30 day mortality (Firth corrected fit):

$$E(Y|\mathbf{L}, C; \boldsymbol{\beta}, \boldsymbol{\psi}) = \text{expit}\left(\mathbf{L}\,\boldsymbol{\beta} + \sum_{c=1}^{m} \psi_c I(C = c)\right)$$

$$\hat{E}(Y(c)) = \frac{1}{n}\sum_{i=1}^{n} \text{expit}\left(\mathbf{L_i}\,\hat{\boldsymbol{\beta}} + \widehat{\psi_c}\right)$$

Acute stroke patients in Sweden

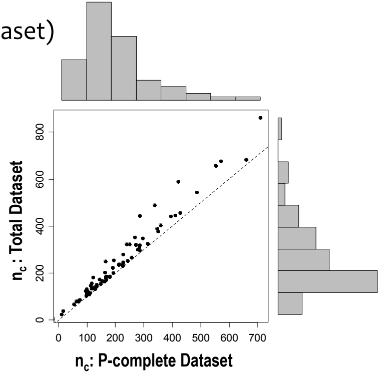
- (MAR) MI vs. CC on the standardized 3 months risk ?
- Dataset explored:
  - $>$ 18 years registered with first stroke in 2011
  - N $=$ 18,850 across 74 hospitals
- Fit (Firth corrected) logistic regression for risk of D3 (DOD3)
- Derive directly standardized risk estimate for each hospital c:

Trade off:
more or more sophisticated confounders vs.
cost of (accurate) registration ,
missing data and measurement error , analysis challenges

RESULTS: Descriptives

- **Number of patients per hospital:**

  - **74 hospitals:** 24 to 861 patients (Total Dataset)
  - 7 hospitals: < 100 patients (Included)

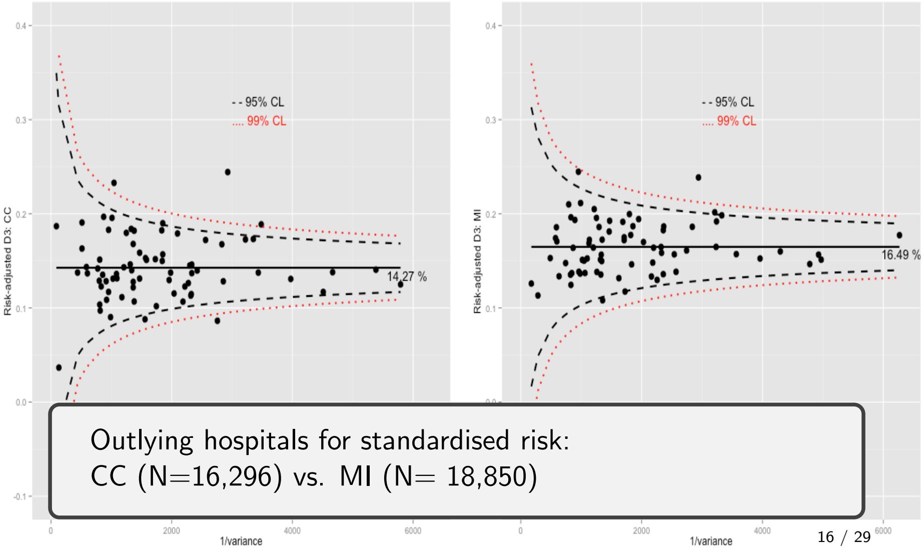


- **Missing predictor variables:**

Number of missing predictor variables	Frequency of patients	Percentage of patients	Cum. frequency	Cum. percentage
0 variables	16307	86.51	16307	86.51
1 variable	2066	10.96	18373	97.47
2 variables	388	2.06	18761	99.53
3 variables	70	0.37	18831	99.90
4 variables	17	0.09	18848	99.99
5 variables	2	0.01	18850	100.00

Only 2.5% patients with 2 or more missing predictor variables

CC (left) and MI (right) and standardized risk



Outlying hospitals for standardised risk:
CC (N=16,296) vs. MI (N= 18,850)

MI vs CC after bench marking standardized risk

Hospital	*CC*	*MI*	*Hospital*	*CC*	*MI*
(# patients; % missing)			(# patients; % missing)		
Hosp. 6 (404; 11.13%)	High*	High*	Hosp. 5 (490; 31.0%)	Low*	OK
Hosp. 7 (457; 6.56%)	High*	OK	Hosp. 9 (237; 9.3%)	OK	Low
Hosp. 25 (247; 8.10%)	High	OK	Hosp. 60 (186; 9.1%)	OK	Low
Hosp. 34 (441; 10.43%)	OK	High	Hosp. 67 (223; 14.3%)	Low	OK
Hosp. 64 (131; 2.29%)	High*	High			

Reducing the covariate set

Consider the covariate subset $\mathbf{L}_{(\mathbf{S})}$ with $\mathbf{S} = (S_1, \ldots, S_p)$

$$S_j = \begin{cases} 1 & \text{if the } j\text{-th covariate is included} \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, p$$

Reducing the covariate set

Consider the covariate subset $\mathbf{L}_{(\mathbf{S})}$ with $\mathbf{S} = (S_1, \ldots, S_p)$

$$S_j = \begin{cases} 1 & \text{if the } j\text{-th covariate is included} \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, p$$

The corresponding main effects regression model for Y is then:

$$E(Y|\mathbf{L}_{(\mathbf{S})}, C; \boldsymbol{\beta}_{(\mathbf{S})}, \boldsymbol{\psi}_{(\mathbf{S})}) = \text{expit}\left(\mathbf{L}_{(\mathbf{S})}\ \boldsymbol{\beta}_{(\mathbf{S})} + \sum_{c=1}^{m} \psi_{c,(\mathbf{S})} I(C = c)\right)$$

and the directly standardized mortality risk:

$$E_{(\mathbf{S})}\{Y(c); \boldsymbol{\beta}, \boldsymbol{\psi}\} = E\left\{E(Y|\mathbf{L}_{(\mathbf{S})}, C = c; \boldsymbol{\beta}_{(\mathbf{S})}, \boldsymbol{\psi}_{(\mathbf{S})})\right\}$$

→ Estimate fixed effects $(\boldsymbol{\beta}_{(\mathbf{S})}, \boldsymbol{\psi}_{(\mathbf{S})})$ with Firth correction:
avoid shrinkage & maintain convergence (Varewyck et al., 2014).

The error functions

Find subset **S** which

- respects the budget $B : \sum_{j=1}^{p} I(S_j = 1) b_j \leq B$, where b_j the j-th covariate cost

The error functions

Find subset **S** which

- respects the budget $B : \sum_{j=1}^{p} I(S_j = 1)b_j \leq B$, where b_j the j-th covariate cost
- Minimizes the error on

1. Error on the predicted individual outcome

$$ER_1(\mathbf{S}) = \left[E\left\{E\left(Y|\mathbf{L}^*_{(\mathbf{S})}, C^*; \hat{\boldsymbol{\beta}}_{(\mathbf{S})}, \hat{\psi}_{(\mathbf{S})}\right) - Y^*\right\}^2\right]^{1/2}$$

- Estimate model parameters $(\boldsymbol{\beta}_{(\mathbf{S})}, \psi_{(\mathbf{S})})$:

  based on 80% of the data $(Y, \mathbf{L}_{(\mathbf{S})}, C)$
- Evaluate error $ER_1(\mathbf{S})$:

  based on 20% new data $(Y^*, \mathbf{L}^*_{(\mathbf{S})}, C^*)$

2. Error on the directly standardized risk for each centre

$$ER_2(\mathbf{S}, c) = \left[E\left(\hat{E}_{(\mathbf{S})}\left\{Y^*(c); \hat{\boldsymbol{\beta}}^*_{(\mathbf{S})}, \hat{\boldsymbol{\psi}}^*_{(\mathbf{S})}\right\} - \hat{E}\left\{Y(c); \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right\}\right)^2\right]^{1/2}$$

- Estimate $(\boldsymbol{\beta}, \boldsymbol{\psi})$ and $\hat{E}\left\{Y(c); \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}\right\}$:

  based on 50% of MI data and all covariates $(Y, \mathbf{L}, C)$
- Estimate model parameters $(\boldsymbol{\beta}^*_{(\mathbf{S})}, \boldsymbol{\psi}^*_{(\mathbf{S})})$ and

$$\hat{E}_{(\mathbf{S})}\left\{Y^*(c); \hat{\boldsymbol{\beta}}^*_{(\mathbf{S})}, \hat{\boldsymbol{\psi}}^*_{(\mathbf{S})}\right\} = \hat{E}\left\{E\left(Y|\mathbf{L}^*_{(\mathbf{S})}, C = c; \hat{\boldsymbol{\beta}}^*_{(\mathbf{S})}, \hat{\boldsymbol{\psi}}^*_{(\mathbf{S})}\right)\right\}$$

based on 50% new (CC or MI) data $(Y^*, \mathbf{L}^*_{(\mathbf{S})}, C^*)$

Selection criterion: $ER_2(\mathbf{S}) = E\{ER_2(\mathbf{S}, c)\}$

Search algorithms

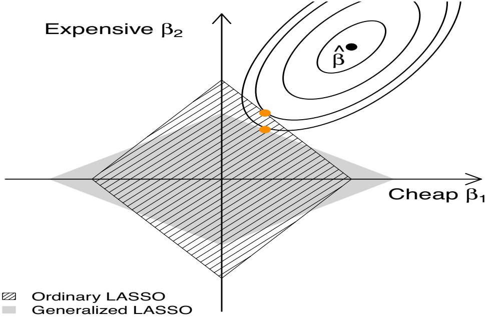
- The parallel hill climber
  - Searches among neighbours in the covariate space
  - that respect the cost constraint
  - for reduced error , improving it with every step
  - 10 parallel chains were used by us
- The parallel tempering algorithm
  - As above but also allows steps that go in the wrong error direction in order to avoid staying in local minima

The generalized LASSO

Investigate the the use of
a weighted penalty function for LASSO regression:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}}\left[\sum_{i=1}^{n}\{Y_i - E(Y_i|\mathbf{L}=\mathbf{L}_i, C=C_i; \boldsymbol{\beta}, \boldsymbol{\psi})\}^2 + \lambda\left(\sum_{j=1}^{p} b_j|\beta_j| + \sum_{c=1}^{m} w_c|\psi_c|\right)\right]$$

where $\lambda \geq 0$ is the tuning parameter



The RAND data (Kahn, 1990)

A sample of $n = 2532$ elderly American patients with pneumonia

- Predict patient's 30-day mortality risk based on subset of $p = 83$ characteristics
- Restrict total covariate cost to 10
- No data were missing & no info on center

	Variable	
Index	Name	Cost
1	Systolic blood pressure score	0.5
2	Age	0.5
3	Blood urea nitrogen	1.5
4	APACHE II coma score	2.5
5	Shortness of breath day 1	1.0
	...	
48	Total APACHE II score	10.0
	...	
83	Sex of patient	0.5

The RAND data

Selection method	Prediction error $\hat{ER}_1(\mathbf{S})$	Total cost (constraint)	Computation time	No. selected covariates
Full model	0.3162	103 (-)	7.3 secs	83
RAND committee	0.3126	30.5 (-)	0.7 secs	14
Population RJMCMC (Fouskakis,2009)	0.3179	10 (10)	3.3 days	8
Parallel hill climber	0.3039	10 (10)	38 mins	13
Parallel tempering	0.3039	10 (10)	2.2 hrs	13
Generalized LASSO with cost constraint	0.3218	9 (10)	9.6 secs	15

→ The stochastic hill climber is preferred selection method here

Swedish register for stroke

Sample of 124 308 patients treated for stroke
in one of 80 Swedish hospitals between 2007 and 2012

- 30-day mortality as quality indicator (never missing)
- 18 baseline patient characteristics (some missing)
- Restrict total allowed cost to 7 ($\sim$ % missing)

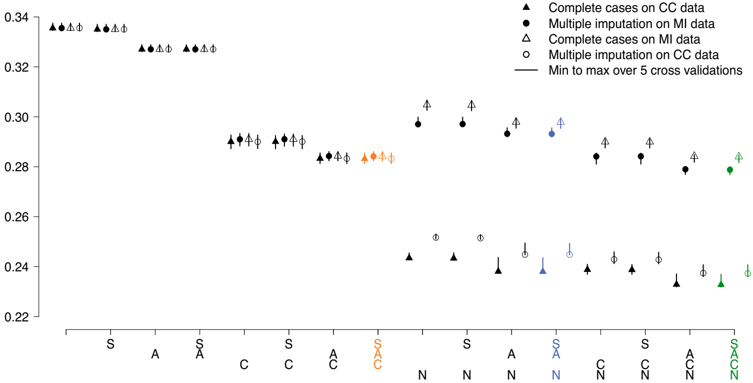
	Descriptive	Missing	Cost	Univariate analysis	
		(%)		Odds ratio	*p*-value
Male	50.9%	0	1	1.40	< 0.001
Age (in years) (Mean & sd)	75.3 (12.4)	0	1	1.06	< 0.001
Consciousness at admission		1.1	1.5		< 0.001
(Alert)	82.6%				
Drowsy	12.1%			8.60	
Unconscious	5.3%			38.71	
NIHSS (Mean & sd)	7.1 (8.8)	66.2	3	1.09	0.018
...					

Minimize the error

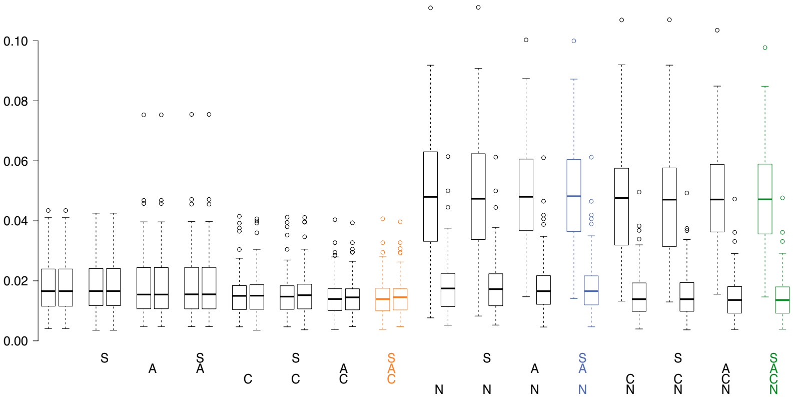
Table : Parallel hill-climber on MI data for Riksstroke

	Patient level	Hospital level
	$ER_1(\mathbf{S})$	$ER_2(\mathbf{S})$
Estimated error	0.2787	0.0161
Cost	6.5	7
Computation time	5.6 hours	7.1 hours
Included covariates	consciousness	consciousness
	NIHSS	NIHSS
	age	
	stroketype	
		year of admission
		patient's ADL-dependence

Prediction errors for individual risk



Prediction errors for standardised risk



Conclusion and discussion

- Enormous methodological progress made
  - causal inference methodology per se
  - general modelling involved (incl. flexible models, robustness, missing data, measurement error)
- impact on routine data analysis limited
- more is often needed in terms of
  - basic interpretation (which question?)
  - assumptions acknowledging, checking
  - transportability: internal versus external validity

STRATOS... and the mission of Malala