FACULTY OF SCIENCES

An overview of developments on Risk Adjustment from the causal inference literature

Stijn Vansteelandt — stijn.vansteelandt@ugent.be
Ghent University, and the London School of Hygiene and Tropical Medicine

GHENT UNIVERSITY

STANDARDISATION

- The evaluation of hospital performance in terms of a quality indicator Y is ideally based on standardisation techniques.

  (Ash et al., 2012; Varewyck et al., 2014)

- Direct standardisation conceptualises for each patient what value Y^c the quality indicator would have taken, had this patient received the levels of care of hospital c.
- It is based on contrasts

$$E(Y^c) \quad \leftrightarrow \quad E(Y^{c^*})$$

- Indirect standardisation evaluates how results for a given center c would have been had the levels of care of a randomly chosen other hospital C^* applied.
- It is based on contrasts

$$E(Y^c|C = c) \quad \leftrightarrow \quad E(Y^{C^*}|C = c)$$

COMPARING APPLES WITH APPLES...

- Such standardisation would be easy if patient case mix was the same in all centers.
- In practice, it is not.
- Standardisation techniques therefore attempt to compare like with like.
- This brings many challenges.

CHALLENGES AHEAD

To adjust for differences in case mix...

- ... we need data on sufficient patient characteristics L,
  so that patients from different hospitals with the same value L can be assumed to be 'alike'.
  - Assume that we have access to data on all prognostic factors of the quality indicator
    that are differentially distributed between hospitals
    and collected upon admission.
- We need models to 'borrow' information between patients.
- E.g., if

$$P(Y = 1 | C = c, L) = \operatorname{expit}(b_c + \beta L) \quad \text{for all } c$$

  then direct standardisation may be based on

$$\hat{E}(Y^c) = \frac{1}{n} \sum_{i=1}^{n} \operatorname{expit}(\hat{b}_c + \hat{\beta} L_i).$$

- Indirect standardisation happens analogously.

Challenges

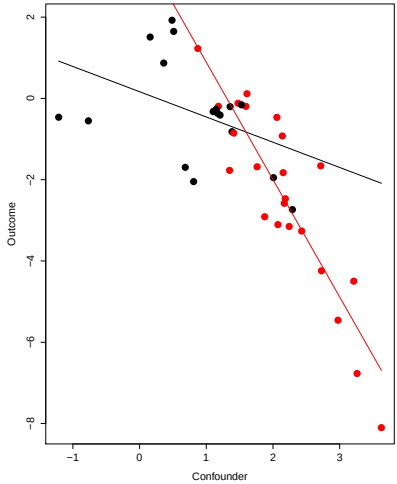
Key questions

- Which patient characteristics to include in those models?
- How to model those patient characteristics?

Our decision how to select and model patient characteristics can be impactful, especially when case mix varies much between hospitals.

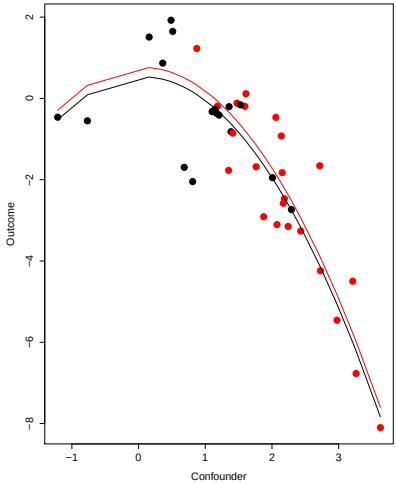
Model misspecification bias

Even well fitting models may then imply serious extrapolations.



MODEL MISSPECIFICATION BIAS

Even well fitting models may then imply serious extrapolations.



Model misspecification bias

- The reason that well fitting models may imply serious extrapolations is that standard fitting strategies are optimised for in-sample prediction.
- These different models deliver nearly the same predictions for the observed data

$$\operatorname{expit}(\hat{b}_{C_i} + \hat{\beta} L_i)$$

- But they yield drastically different predictions away from the observed data, and thus different estimators

$$\frac{1}{n} \sum_{i=1}^{n} \operatorname{expit}(\hat{b}_c + \hat{\beta} L_i).$$

- To prevent this, the statistical analysis should be tuned towards the study's aim.

SHRINKAGE BIAS

- Bias also arises when shrinkage estimators $\hat{b}_c$ are used, e.g., obtained by fitting random effects models.
- Because bias is towards the 'center', it reduces power.

  (Ash et al., 2012)
- In previous work, we have aimed for a compromise using Firth correction, with reasonably good success.

  (Varewyck et al., 2014)
- This correction has also been criticised.

  (Greenland and Mansournia, 2015)
- Previous approaches are designed to minimise mean squared error or bias in $\hat{b}_c$, but not in $\hat{E}(Y^c)$.
- Regardless of the penalisation method used, shrinkage bias remains a concern.

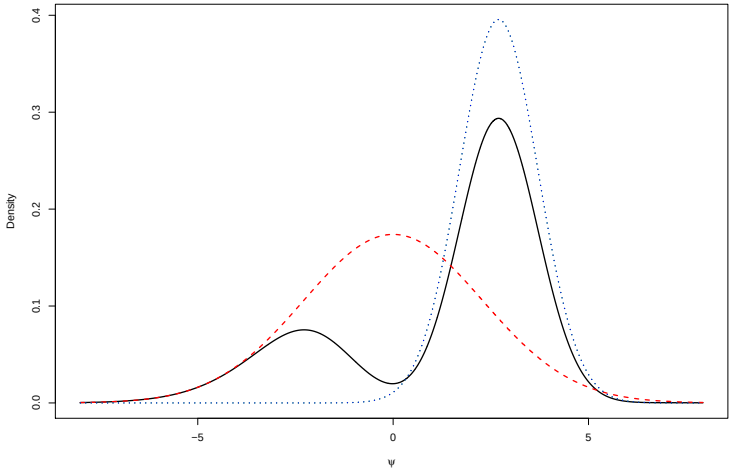
VARIABLE SELECTION BIAS

- Selecting the ‘wrong’ variables for adjustment may leave residual confounding bias.
- Bias can be large.
- E.g., moderate prognostic factors can be difficult to detect
  when they are highly differential between hospitals.
- Standard selection strategies are designed to give good in-sample predictions.
- They are not optimised for standardisation.
- This not only biases estimators, but also uncertainty assessments
  as a result of ignoring uncertainty in the choice of variables / model.

(e.g. Leeb and Pötscher, 2006; Dukes and Vansteelandt, 2020)

VARIABLE SELECTION BIAS

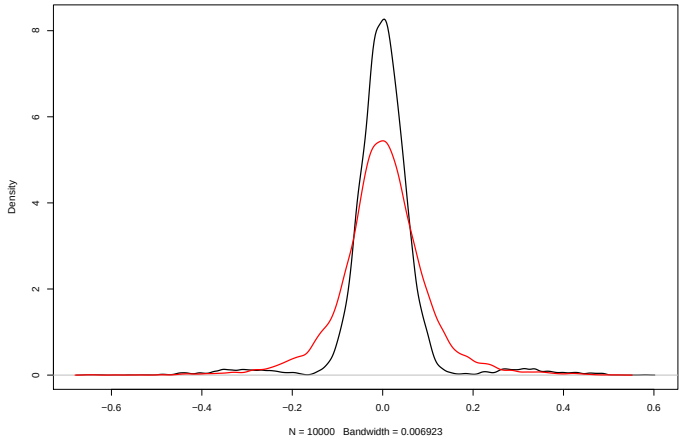
Our ‘hesitation’ for which components of L to adjust,
creates bias and inefficiency, and invalidates standard inference.



Empirical (black) versus bootstrap (red) distribution

It may be tempting to consider the bootstrap,
but this has no theoretical justification.

(e.g. Samworth, 2011)



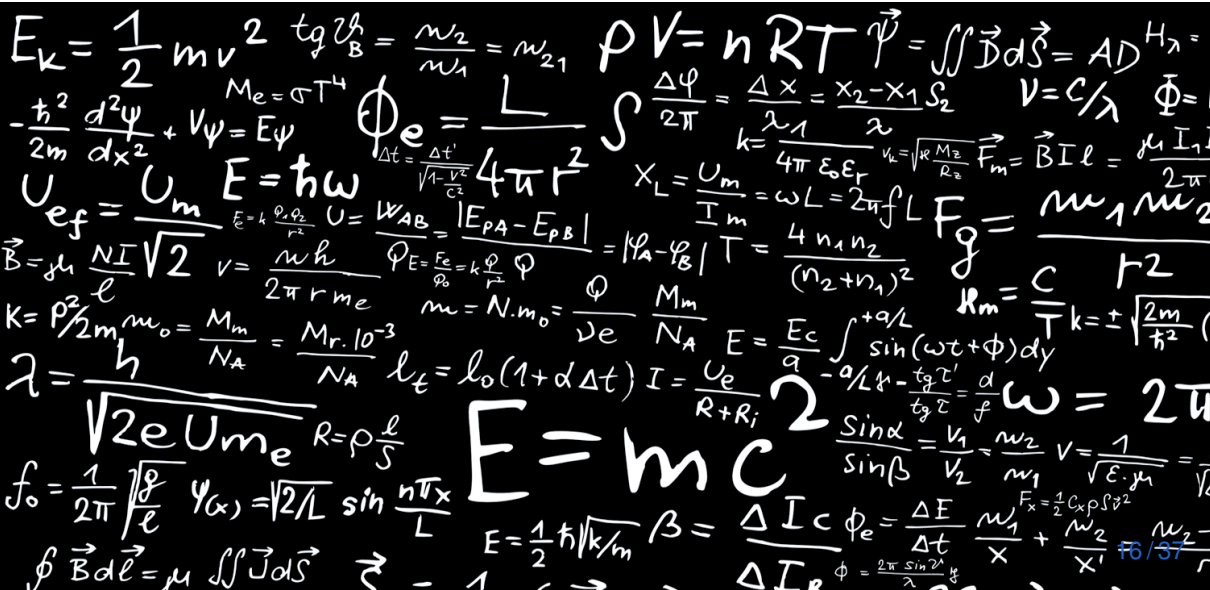
SUMMARY SO FAR: STANDARD METHODS ARE POORLY TUNED

- Routine standardisation approaches can be sensitive to bias due to model misspecification, shrinkage and variable/model selection.
- This is partly the result of poor tuning of standard statistical methods.
- It may imply especially severe bias when information is lacking, e.g., when the outcome is rare or there are large differences in case mix.

SUMMARY SO FAR: STANDARD METHODS ARE DIFFICULT TO PRE-SPECIFY

- Model building is needed to prevent overfitting and model misspecification.
- Standard model building approaches are difficult to pre-specify.
- This can make the analysis prone to error.
- Standard confidence intervals ignore uncertainty about the model that will be fitted.
- This is a severe problem in nearly all statistical analyses
  as remedies are not well known, even amongst professional statisticians.

SOLUTIONS



DOUBLE ROBUST STANDARDISATION

- Enormous progress on standardisation has been made in the causal inference literature,

  (Hernan and Robins, 2020)

  much of this being centred around double robust standardisation.

  (Robins and Rotnitzky, 2001; Vansteelandt and Keiding, 2011; Varewyck et al., 2014; Rotnitzky and Vansteelandt, 2018; Seaman and Vansteelandt, 2018)

- To infer $E(Y^1)$, this invokes 2 predictions:
  - a propensity score, e.g.
  $$P(C = 1|L) = \text{expit}(\alpha L)$$
  - and an outcome mean, e.g.
  $$E(Y|C = 1, L) = \text{expit}(b_1 + \beta L)$$

- I will first explain how this works in the absence of shrinkage or variable selection.
- I will next give an improvement designed to remove shrinkage and variable selection bias.

DOUBLE ROBUST STANDARDISATION FOR $E(Y^1)$

- Ideally, we would like to fit a model for Y^1 based on all patients.
- But we only observe Y^1 for patients in center 1.
- Suppose we knew that only 10% of severely ill patients were in center 1.
- Then by counting those 10 times in the analysis,
  we mimic severely ill patients from the entire sample.
- We will therefore weigh each patient's data before fitting a prediction model,
  to make the patients in center 1 representative of all patients.
- This tunes the model fitting better to the eventual goal.

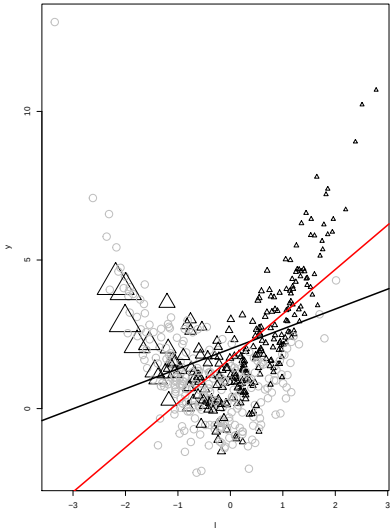
Double robust standardisation for $E(Y^1)$

Formally, we

- estimate $P(C = 1|L)$ using logistic regression as $g(L)$;
  - The estimated propensity score teaches us where predictions will be made.
- estimate $E(Y|C = 1, L)$ in patients of hospital 1 as $Q(L)$ using logistic regression, weighting by $1/g(L)$;
- use this to predict outcome for all;
- average these predictions:

$$\frac{1}{n}\sum_{i=1}^{n} Q(L_i)$$

DOUBLE ROBUST STANDARDISATION FOR $E(Y^1)$



What does it deliver?

- Through better tuning, we reduce model misspecification bias.
- Even when the outcome model is misspecified,
  unbiased predictions are then still obtained if the propensity score is correct.
- And vice versa. Hence the name double robust.
- This is especially relevant with rare outcomes.
- Correct modelling of the outcome may then not be possible in view of overfitting.
- Propensity score methods then tend to be less biased.

  (Joffe and Rosenbaum, 1999; Braitman and Rosenbaum, 2002; Patorno et al., 2014; Cepeda et al., 2003)
- The main benefits of double robust standardisation are seen
  when additionally considering bias due to shrinkage and variable selection.

A NAIVE ESTIMATOR

- Let's bring in model/variable selection.
- Let $Q^{(0)}(L)$ be an initial prediction of Y^1.
- E.g., based on fitting model

$$E(Y|C = c, L) = \text{expit}(b_c + \beta L) \quad \text{for all } c$$

  using logistic mixed effects model after adopting a stepwise variable selection procedure.
- Direct standardisation delivers an estimator

$$\frac{1}{n}\sum_{i=1}^{n} Q^{(0)}(L_i)$$

  subject to shrinkage and variable selection bias.

TUNING THE OUTCOME PREDICTIONS

- To tune the outcome predictions, we need a propensity score estimator $g(L)$.
- E.g. based on fitting model

$$P(C = 1|L) = \text{expit}(\alpha L)$$

  using logistic regression after adopting a stepwise variable selection procedure.
- Running a weighted regression for the outcome helps,
  but is not generally good enough in these more complex settings.

TARGETED/CAUSAL LEARNING

- We instead remove shrinkage bias by basing direct standardisation on the logistic model

$$E(Y|C=1,L) = \operatorname{expit}(V + \delta W)$$

in hospital $C = 1$ for offset

$$V = \operatorname{logit} Q^{(0)}(L)$$

and covariate

$$W = \frac{1}{g(L)}$$

- Let the fitted values (based on the MLE $\hat{\delta}$ for δ) be $Q^{(1)}(L) = \operatorname{expit}(V + \hat{\delta} W)$.
- We then report the estimator

$$\frac{1}{n}\sum_{i=1}^{n} Q^{(1)}(L_i)$$

(van der Laan and Rubin, 2006; van der Laan and Rose, 2014)

WHY TARGETED/CAUSAL LEARNING?

- This estimator is also double robust.
- When both models are correctly specified,
  then its behaviour is not sensitive to shrinkage or variable selection bias.

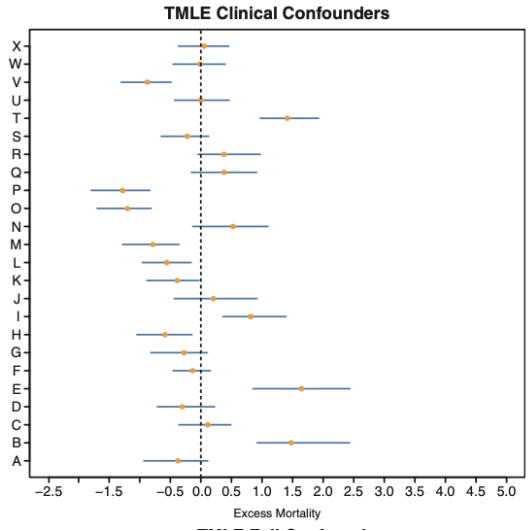
  (van der Laan and Rubin, 2006; van der Laan and Rose, 2014)
- The implication is that valid confidence intervals can be obtained via the bootstrap,
  or even simple analytical calculations.
- At a more intuitive level, this is because confounders now have 2 chances of being detected:
  once in the propensity score model, and once in the outcome model.
- Technically, this is because the estimator's bias is the product of bias in both models,
  and this makes shrinkage biases become negligible.
- Again, this is especially relevant with rare outcomes.

TARGETED LEARNING FOR HOSPITAL PROFILING

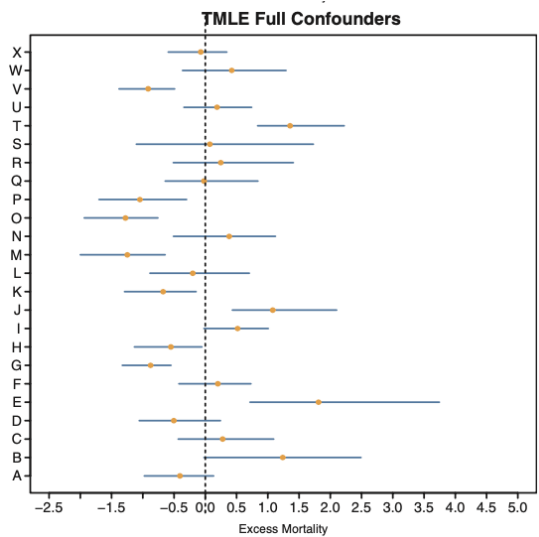
- Hospital all-cause 30-day excess mortality risk among 8952 adults undergoing percutaneous coronary intervention between Oct 1, 2011, and Sept 30, 2012.
- Clinical registry data from 24 Massachusetts hospitals, linked with billing data.

  (Spertus et al., 2016)
- Variable selection and shrinkage among 225 confounders using ϵ-net procedure.

TARGETED LEARNING RESULTS ($p = 11, n = 8952/24$)



TARGETED LEARNING RESULTS ($p = 225, n = 8952/24$)



What if models are wrong?

- These results are extremely powerful.
- But demand both models to be correctly specified.
- What if they are not?

CAUSAL MACHINE LEARNING

- One may lessen the risk of model misspecification by using (machine) learning algorithms to estimate propensity score and obtain (initial) outcome predictions.
- Machine learning must be viewed in the broad sense as any (combination of) procedure(s) that is data-adaptive and can be automated.
- van der Laan proposed this in his theory on Targeted Maximum Likelihood Estimation or Targeted Learning.

  (van der Laan and Rubin, 2006; van der Laan and Rose, 2014)
- Available in user-friendly R-package `tmle`.
- Related proposals were recently made by Belloni, Chernozhukov, Newey, Robins, ...

  (Chernozhukov et al., 2018)
- They refer to their approach as double machine learning.

AN IMPRESSION...

- For just 2 hospitals $C = 0, 1$, consider estimating $E(Y^1) - E(Y^0) = 0.5$ in model

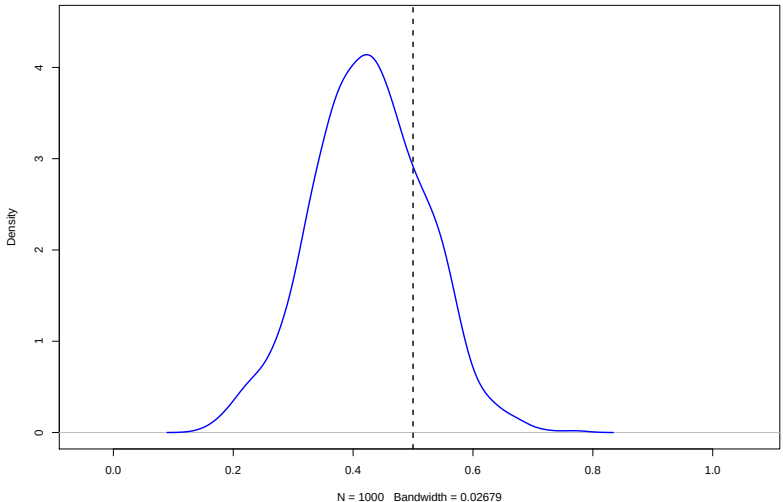
$$E(Y|C, L) = \psi C + 1.25 \cos^2(\gamma' L)$$

with standard normal noise and

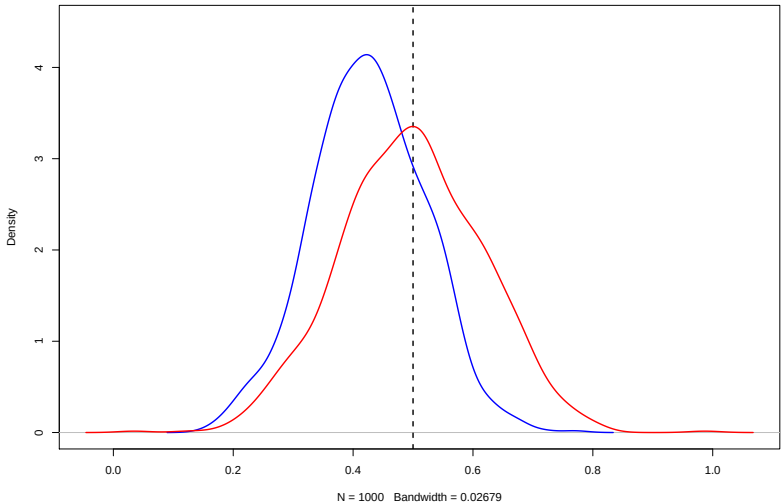
$$P(C = 1|L) = \text{expit}\left\{2 \cos(\gamma' L) + 2 \sin(\gamma' L)\right\}.$$

- $p = 10$, $n = 500$.
- I will contrast direct standardisation based on random forest regression with a targeted learning procedure based on it.

NAÏVE MACHINE LEARNING USING RANDOM FORESTS



CAUSAL MACHINE LEARNING USING RANDOM FORESTS



Conclusions from the simulation study

- Targeted learning removes bias from improper tuning of random forest regression.
- Valid standard errors were obtained,
  despite the uncertainty of random forests predictions not being known.

SUMMARY

- Double robust standardisation tunes model fitting strategies towards the eventual aim.
- This reduces model misspecification bias.
- With minor modifications, this can incorporate flexible model building algorithms with the advantage that it
  - reduces shrinkage and variable selection bias;
  - delivers valid confidence intervals.
- It moreover allows one to incorporate (machine) learning algorithms, so that the analysis can be pre-specified.
- The resulting approach is known as targeted / causal learning.

IMPROVEMENTS

- Targeted learning delivers efficient estimators and valid confidence intervals
  when center and outcome are correctly modelled,
  but not otherwise.
- When both are incorrectly modelled,
  they can sometimes be sensitive to bias.

  (Kang and Schafer, 2007)
- This may not be an enormous concern when flexible methods are used,
  except when the information is scarce.

IMPROVEMENTS

- To handle these situations, vigorous research activity has lead to drastic improvements,

  (reviewed in Rotnitzky and Vansteelandt, 2018)

  mostly focussed on parametric models with variable selection,
  though not exclusively.

  (Benkeser et al., 2017; Dukes, Whitney and Vansteelandt, 2021)

- Empirical efficiency maximisation is designed to deliver efficient estimators
  when one model is misspecified.

  (Rubin and van der Laan, 2008; Cao, Tsiatis and Davidian, 2009)

- Bias-reduced double robust estimation is designed
  to prevent bias amplification when both models are wrong,
  and delivers valid confidence intervals even when one model is wrong.

  (Vermeulen and Vansteelandt, 2015; Avagyan and Vansteelandt, 2021)