Causal (in)validity of the trimmed means estimand in clinical trials

Jonathan Bartlett
www.thestatsgeek.com

Department of Mathematical Sciences
University of Bath, UK

15th July 2022

Acknowledgements / declarations

This is joint work with Camila Olarte Parra (Bath) and Rhian Daniel (Cardiff).

This is work in progress. Any mistakes are mine!

This work was supported by a UK Medical Research Council grant (MR/T023953/1).

My institution has received consultancy income for my advice on statistical methodology from AstraZeneca, Bayer, Novartis, Roche and I have received consultancy income from Bayer and Roche.

ICH E9 estimand addendum

Trimmed means estimand

What is a causal effect/estimand anyway?!

Conclusions

Outline

ICH E9 estimand addendum

In 2019 ICH published 'E9 (R1) addendum on estimands and sensitivity analysis in clinical trials'

It describes framework for defining clinical trial estimands

Estimand requires (according to this) specification of 5 attributes:

- the **treatments** being compared
- the **population** of patients targeted
- the **variable** to be obtained on each patient
- the strategies to handle **intercurrent events**
- the **population summary measure**, used to compare treatment groups

Intercurrent events

Intercurrent events (ICEs) are defined as:

> *events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest*

e.g. discontinuation or randomised treatment due to lack of efficacy or toxicity.

ICH E9 intercurrent event strategies

- treatment policy
- hypothetical
- **composite - incorporate occurrence of intercurrent event in endpoint/variable definition**
- while on treatment
- principal stratification

Outline

Intercurrent events as 'failures'

- RCT with continuous outcome Y.
- Intercurrent event e.g. discontinuation of randomised treatment for lack of efficacy or toxicity, represents 'failure'.
- We take a composite type approach and assign 'bad' outcome value to patients experiencing this.
- But what value to assign?

Trimmed means estimand

- Permutt and Li (2017) proposed an approach where we assign an arbitrarily low/high bad value.
- Trim (remove) the worst α% of outcome values from each treatment group.
- Calculate difference in 'trimmed means' between treatment groups.
- Choice of α needs to be sufficiently large that all patients with intercurrent event are trimmed (removed).

How to interpret trimmed means estimand?

On the interpretation of the resulting estimand, Permutt and Li (2017) write:

> *Some patients did badly on treatment. Either they completed with bad outcomes, or they dropped out. For some medical conditions, it will not matter much whether they dropped out or completed with bad outcomes, nor how bad the bad outcomes were. The trimmed mean is the average outcome for other patients, those who did best in each group.*

But is a contrast of trimmed means a causally valid estimand?

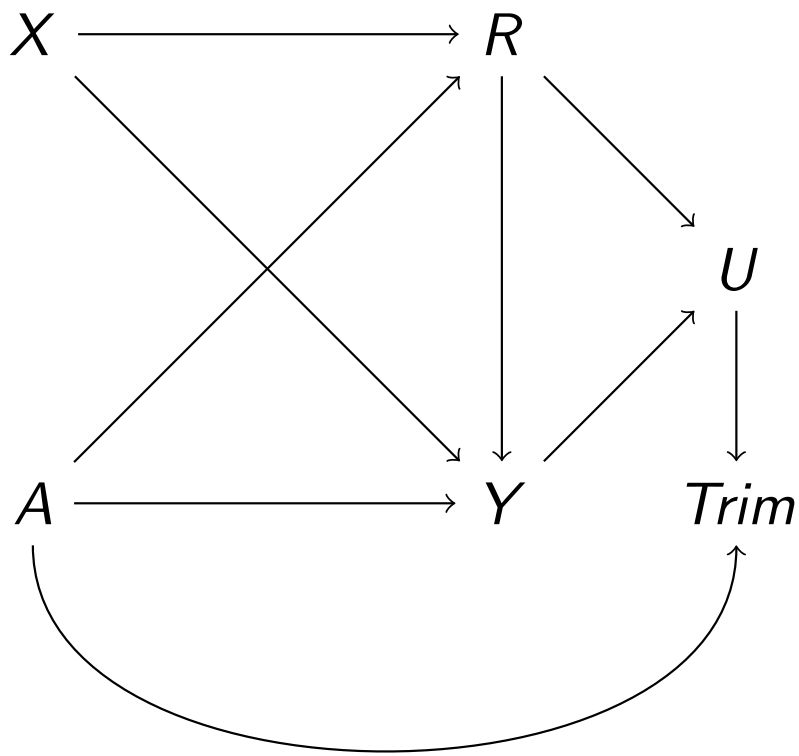
Some notation

- Randomised treatment A.
- Continuous outcome Y.
- Baseline variables X which affect outcome Y (some will be unmeasured).
- Indicator of intercurrent event **not occurring** R ($R = 0$ if occurred, $R = 1$ if not).
- Composite outcome U:

$$U = \begin{cases} -\infty & \text{if } R = 0 \\ Y & \text{if } R = 1 \end{cases}$$

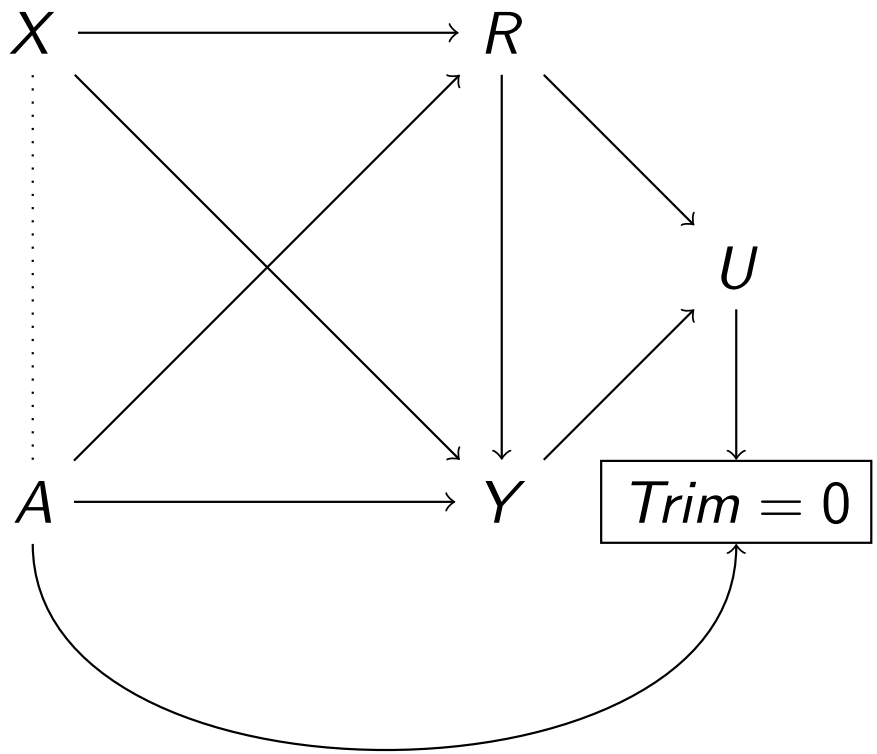
- Indicator of being trimmed $Trim$: $Trim = 1(U < F^{-1}_{U^A}(\alpha))$.
- $F^{-1}_{U^a}(\alpha))$ is the α quantile of U^a, where U^a is the potential outcome for U under assignment to treatment a.

Directed acyclic graph (DAG)



DAG after 'trimming'

Analysing after trimming corresponds to conditioning on $Trim = 0$:



Among the remaining patients, treatment groups (A) are no longer balanced with respect to baseline variables X (indicated by dotted line).

Implications

- Groups being compared in trimmed means are not exchangeable.
- Comparisons of trimmed means between treatment groups are generally confounded by differences in baseline variables.
- We often measure and adjust (e.g. using regression models) for some baseline covariates.
- But there will always exist some we do not measure, and if these exist, they may lead to the trimmed means estimand being confounded.

Results for a simple setup

- Suppose X is a single binary variable, with $P(X = 1) = \pi$.
- Let $\delta_{ax} = P(R = 1|A = a, X = x)$, $x = 0, 1$, $a = 0, 1$.
- Let $\tau_{ax}(.)$ denote the CDF of Y among those with $R = 1$, $A = a$ and $X = x$.
- Then one can show

$$P(X = 1|A = a, Trim = 1) = \frac{\pi\delta_{a1}\left[1 - \tau_{a1}(F^{-1}_{U(a)}(\alpha))\right]}{1 - \alpha}$$

- The proportion with $X = 1$ among those not trimmed will generally differ between $A = 0$ and $A = 1$.

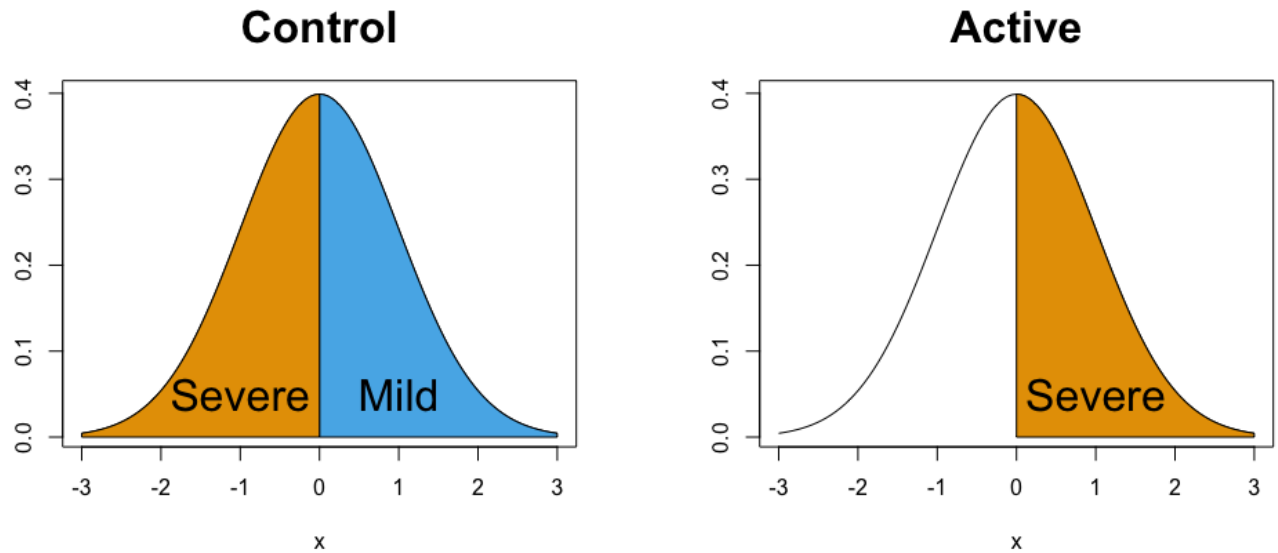
Results for a simple setup

- Suppose in control arm 10% experience ICE among those with $X = 0$ ($\delta_{00} = 0.9$) and 20% experience ICE among those with $X = 1$ ($\delta_{01} = 0.8$).
- Active treatment reduces ICE, with 5% occurrence in those with $X = 0$ ($\delta_{10} = 0.95$), and 10% in those with $X = 1$ ($\delta_{11} = 0.9$).
- Suppose $Y|A, X, R = 1 \sim N(\beta_0 + \beta_1 A + X, 1)$, and we choose $\alpha = 0.3$.
- Then
  - $P(X = 1|A = 0, Trim = 0) = 0.54$
  - $P(X = 1|A = 1, Trim = 0) = 0.58$

Results for a contrived but illuminating setup

- Consider this extreme but hopefully helpful setup, inspired by one in Permutt and Li (2017).
- Suppose population consists of mild & severe patients, in 1:1 ratio.
- On control, outcomes Y^0 are $N(0, 1)$, with higher values being better.
- Mild patients are the positive half normal part, and severe patients are the negative half normal part.
- Suppose no intercurrent events occur on control.
- On active, patient has intercurrent event if and only if they are a mild patient.
- For severe patients, who don't have intercurrent event, suppose $Y^1 = -Y^0$.
- Suppose we trim $\alpha = 0.5$ from each treatment group.

Results for a contrived but illuminating setup



If we trim the worst (lowest) 50%, we are comparing mild patients to severe patients when we compare treatment groups.

Outline

What is a causal effect/estimand anyway?!

- What constitutes a causally valid effect measure/estimand?
- From Section A3 of ICH E9 addendum:
  An estimand is a precise description of the treatment effect reflecting the clinical question posed by a given clinical trial objective. It summarises at a population level what the outcomes would be in the same patients *under different treatment conditions being compared.*
- By this definition, the trimmed means estimand fails - it is not comparing the same patients under two treatment conditions.

What is a causal effect/estimand anyway?!

- Hernán and Robins (2020) state
  a population causal effect may also be defined as a contrast of functionals, including medians, variances, hazards, or cdfs of counterfactual outcomes
- Based on this Ocampo et al. (2021) observe that the trimmed means estimand satisfies this, since it is equal to

$$T_\alpha(F_{U^1}) - T_\alpha(F_{U^0})$$

  where

$$T_\alpha(F) = \frac{1}{1-\alpha}\int_{F^{-1}(\alpha)}^{\infty} y \; dF(y)$$

- They note though that it may be challenging for clinicians and patients to interpret.

Are hazard ratios valid causal effects?

- By the same definition, the hazard ratio at time $t > 0$ in an RCT comparing treatment groups is a valid causal effect.
- This is because

$$HR(t) = \frac{f^0(t)/S^0(t)}{f^1(t)/S^1(t)}$$

  where $f^0(t)$ and $f^1(t)$ are the population/marginal densities of the counterfactual failure times and $S^0(t)$ and $S^1(t)$ the corresponding survival functions
- Despite this, many have criticised the hazard ratio in terms of its causal interpretability (e.g. Stensrud et al. (2019)), since $HR(t)$ compares event rates among survivors to time t between treatment groups, and in general these patients are not exchangeable.

Outline

Conclusions

- Interpreting the trimmed means *estimand* causally is tricky.
- The best $(1-\alpha)\%$ patients if you assign the population to receive control is not generally the same best group of patients if you assign the population to receive active.
- It is debatable whether a trimmed mean is a summary measure of the whole group/population or not.
- Note the trimmed means *estimator* can instead be viewed as targeting a full population average treatment effect estimand under certain assumptions (Ocampo et al., 2021; Hazewinkel et al., 2022).

References

Hazewinkel, A.-D., Bowden, J., Wade, K. H., Palmer, T., Wiles, N. J., and Tilling, K. (2022). Sensitivity to missing not at random dropout in clinical trials: use and interpretation of the trimmed means estimator. *Statistics in Medicine*, 41(8):1462–1481.

Hernán, M. and Robins, J. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Ocampo, A., Schmidli, H., Quarg, P., Callegari, F., and Pagano, M. (2021). Identifying treatment effects using trimmed means when data are missing not at random. *Pharmaceutical statistics*, 20(6):1265–1277.

Permutt, T. and Li, F. (2017). Trimmed means for symptom trials with dropouts. *Pharmaceutical Statistics*, 16(1):20–28.

Stensrud, M. J., Aalen, J. M., Aalen, O. O., and Valberg, M. (2019). Limitations of hazard ratios in clinical trials. *European heart journal*, 40(17):1378–1383.