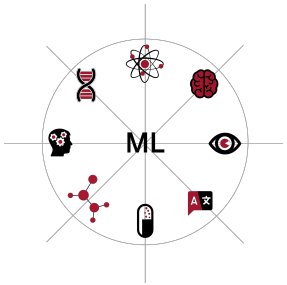
Can Unbiased Estimation Justify Randomized Trials?

Konstantin Genin

Group Leader: "Epistemology and Ethics of ML"

In collaboration with
Conor Mayo-Wilson, University of Washington (Seattle)



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Critics of Randomization

Randomization has come in for criticism on purely **epistemic** grounds.

- Bayesians have a hard time rationally reconstructing randomization (Savage 1961,1962; Kadane & Seidenfeld, 1999; Kasy 2016).

- The frequentist theory of optimal design of experiments does not endorse randomization (Kiefer 1959; Harville 1975).

- Philosophers of science have criticized the coherence of randomization (Urbach 1985; Worrall 2002).

Randomization On its Own Terms

What is the best frequentist justification for randomization?

A standard answer: Randomization guarantees the existence of an unbiased estimator of average treatment effect.

Our critique

1. Non-trivial randomization is **not necessary** for unbiased estimation of ATE.
2. Not sufficient for unbiased estimation of other quantities e.g., the median or minimum causal effect. (No design guarantees unbiased estimation)
3. Unbiasedness is not always a good to be sought:
    a. Unbiased estimates may be incompatible with what is known from data or what is assumed by the model.
    b. Unbiased estimators are often ***inadmissible*** (i.e., weakly dominated) not matter one's loss function.

Our proposal

Minimax foundations demonstrating that randomized designs (which ones?) minimize worst case loss (in what sense?) at **finite** samples.

Our critique

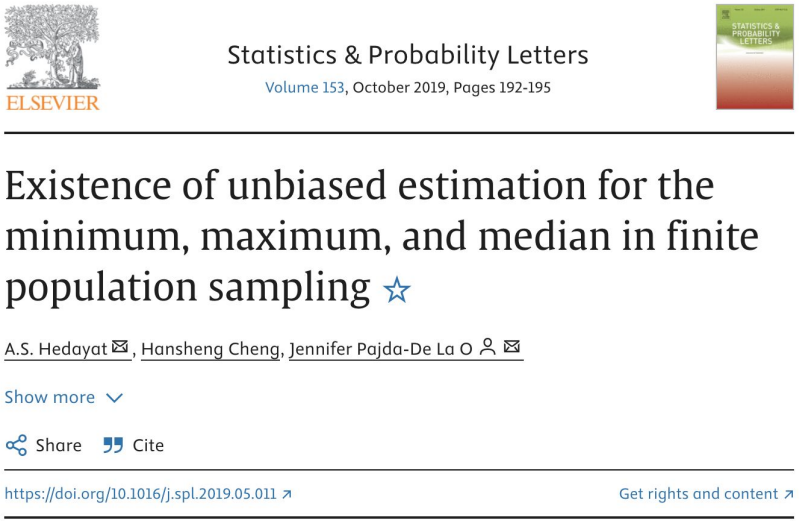
1. Non-trivial randomization is **not necessary** for unbiased estimation of ATE.
2. Not sufficient for unbiased estimation of other quantities e.g., the median or minimum causal effect. (No design guarantees unbiased estimation)
3. Unbiasedness is not always a good to be sought:
    a. Unbiased estimates may be incompatible with what is known from data or what is assumed by the model.
    b. Unbiased estimators are often ***inadmissible*** (i.e., weakly dominated) not matter one's loss function.

Our critique

1. Non-trivial randomization is **not necessary** for unbiased estimation of ATE.
2. Not sufficient for unbiased estimation of other quantities e.g., the median or minimum causal effect. (No design guarantees unbiased estimation)
3. Unbiasedness is not always a good to be sought:
    a. Unbiased estimates may be incompatible with what is known from data or what is assumed by the model.
    b. Unbiased estimators are often ***inadmissible*** (i.e., weakly dominated) not matter one's loss function.

Absence of Unbiased Estimates for Other Quantities

"there are no unbiased estimators of the minimum, maximum, or median for finite population sampling under any sampling design except census" (Hedayat et al., 2019).



ELSEVIER

Statistics & Probability Letters

Volume 153, October 2019, Pages 192-195

Existence of unbiased estimation for the minimum, maximum, and median in finite population sampling ☆

A.S. Hedayat ✉, Hansheng Cheng, Jennifer Pajda-De La O ✉

Show more

Share Cite

https://doi.org/10.1016/j.spl.2019.05.011 ↗ Get rights and content ↗

Absence of Unbiased Estimates for Other Quantities

Note: If the outcome variable is only ordinal-scaled, comparisons of the ATEs of different interventions depend on the measurement unit.

E.g. scholarships could appear better than counseling if you measure the number of college courses that are completed, but they might appear worse if you measure number of semesters completed.

Absence of Unbiased Estimates for Other Quantities

Theorem (Conor).

There are no unbiased estimators for the p^{th} *percentile,* which is the least value x such that at least p percent of the population is below x.

Our critique

1. Non-trivial randomization is **not necessary** for unbiased estimation of ATE.
2. Not sufficient for unbiased estimation of other quantities e.g., the median or minimum causal effect. (No design guarantees unbiased estimation)
3. Unbiasedness is not always a good to be sought:
    a. Unbiased estimates may be incompatible with what is known from data or what is assumed by the model.
    b. Unbiased estimators are often ***inadmissible*** (i.e., weakly dominated) not matter one's loss function.

Inverse Propensity Score Weighting

Suppose you want to estimate patient *i*'s response to treatment $Y_i(1)$. An unbiased estimator is given by:

$$\hat{\theta}(Y = y, T = t) = \begin{cases} 1/p \cdot y & \text{if } t = 1 \\ 0 & \text{if } t = 0 \end{cases}$$

Suppose p = 1/2. As things turn out, patient i is treated (T = 1) and survives four months past treatment. So it is **known** that $Y_i(1) = 4$. But the estimate for $Y_i(1)$ is eight months, i.e., that we estimate that, if she were treated (which she was!), then patient i would survive twice as long as she actually did. *Unbiased estimates may be incompatible with observation!*

Inverse Propensity Score Weighting

This is a special case of a general phenomenon: when parameter spaces are bounded, unbiased estimation requires over(under)shooting *known bounds.* Unbiased estimators are dominated by estimators that cut-off at the known bounds.

“inadmissibility of unbiased estimators is likely to be the rule, rather than the exception” (Berger, 1989).

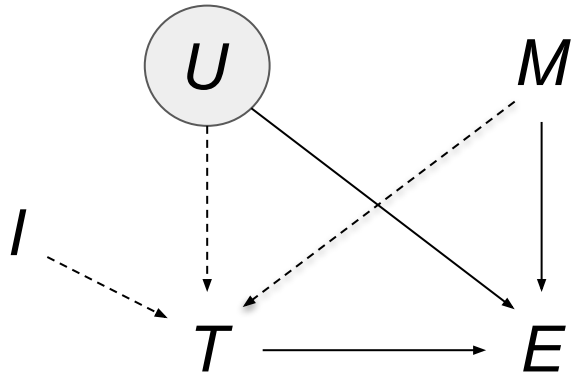
Berger, James O. (1989) “On the Inadmissibility of Unbiased Estimators.” *Statistics and Probability Letters.* 9(5): pp. 381-4.

Our critique

1. Non-trivial randomization is **not necessary** for unbiased estimation of ATE.
2. Not sufficient for unbiased estimation of other quantities e.g., the median or minimum causal effect. (No design guarantees unbiased estimation)
3. Unbiasedness is not always a good to be sought:
    a. Unbiased estimates may be incompatible with what is known from data or what is assumed by the model.
    b. Unbiased estimators are often ***inadmissible*** (i.e., weakly dominated) not matter one's loss function.

The Causal Situation

T := treatment (binary);
E := effect (binary);
M := measured covariates;
U := unmeasured covariates;
I := randomizer.



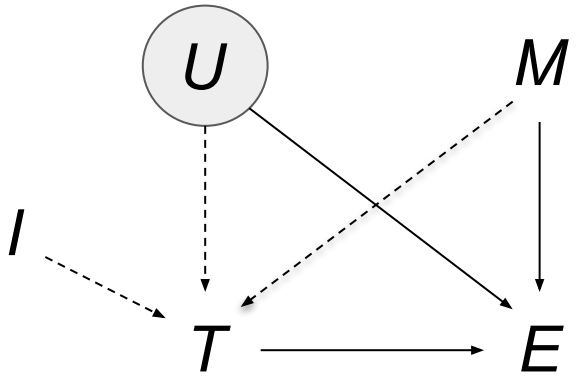
Average Treatment Effect

The goal is to estimate the
average treatment effect (ATE):

$$P(E = 1|do(T = 1)) - P(E = 1|do(T = 0))$$

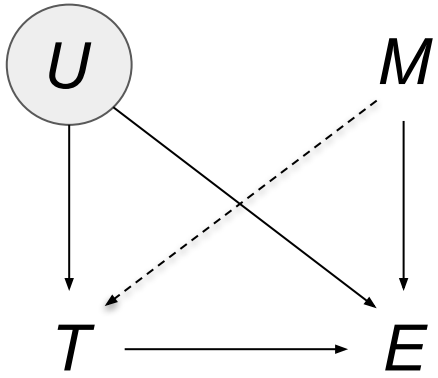
Or, in the notation of the potential outcomes framework:

$$\frac{1}{n}\sum_{i \leq n} P(E_i^{t=1} = 1) - P(E_i^{t=0} = 1)$$



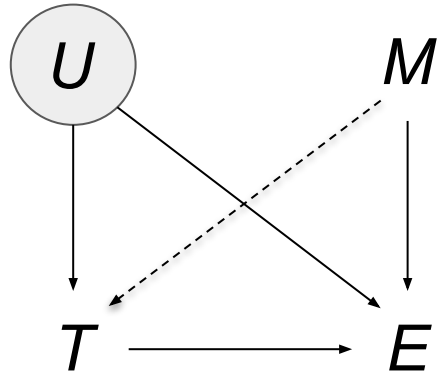
Trouble with Observational Studies

If there is an unobserved common cause of *T, E*
it is easy to come up with examples in which the ATE
is **not identified.**

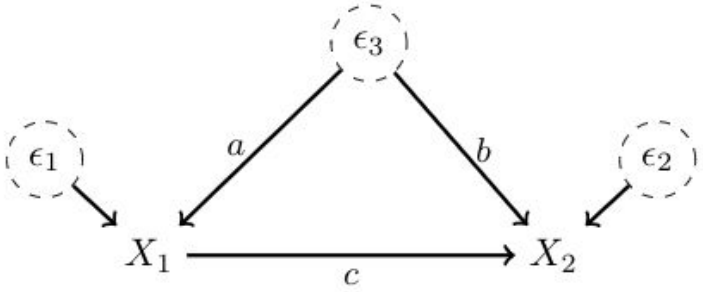


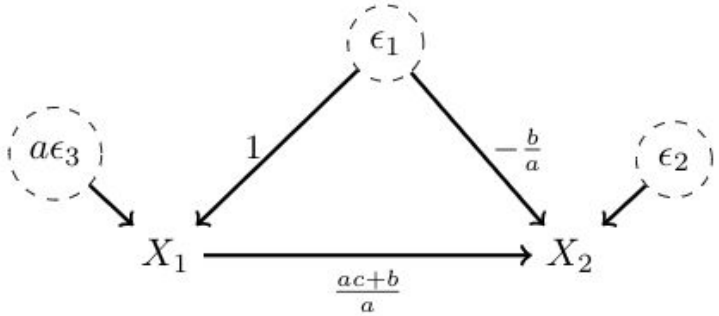
Trouble with Observational Studies

If there is an unobserved common cause of *T, E*
it is easy to come up with examples in which the ATE
is **not identified.**



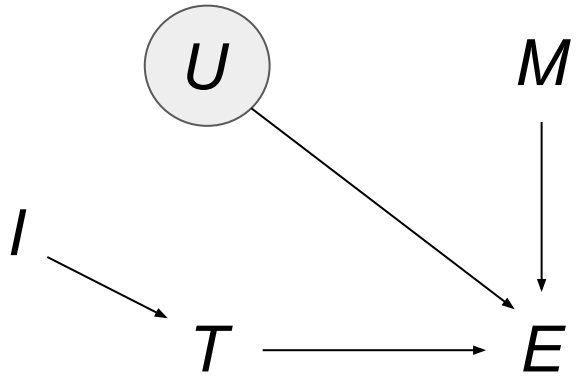
Trouble with Observational Studies





The Point of Randomization

Randomization “breaks edges” into treatment, so that any association between *T* and *E* is due to the causal effect of *T* on *E* and not shared common causes.

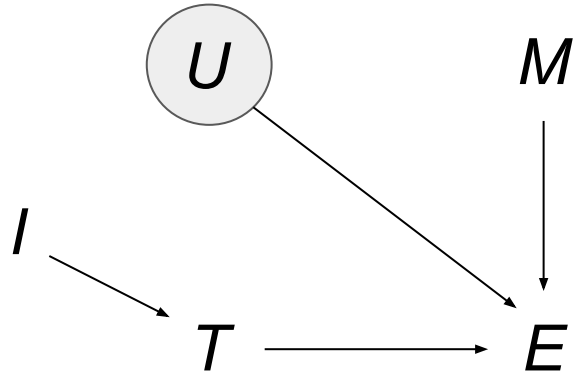


The Point of Randomization

It ensures that the ATE is **identified** and equal to

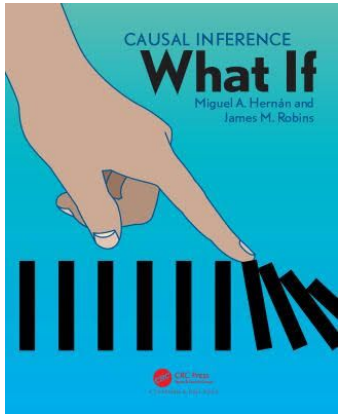
$$P(E = 1|T = 1) - P(E = 1|T = 0)$$

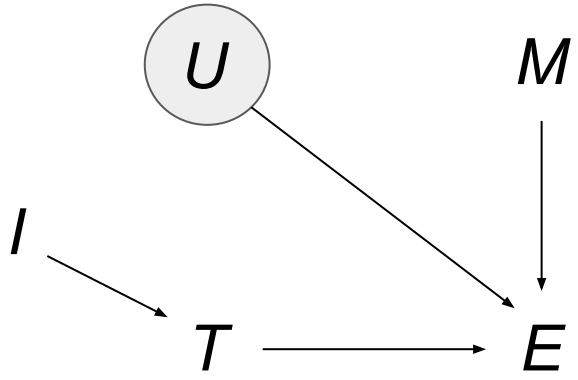
Moreover an **unbiased estimate** of the ATE is easily obtained.



The Point of Randomization

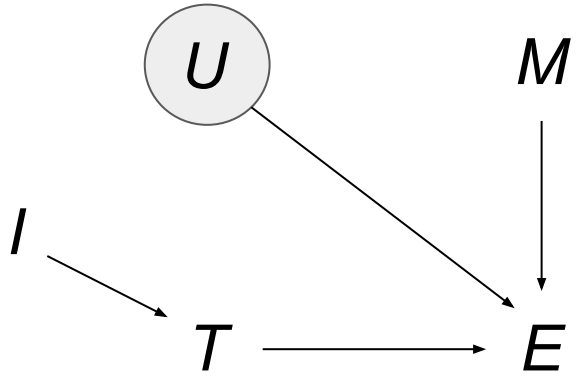
“In ideal randomized experiments, association is causation”





No Other Way?

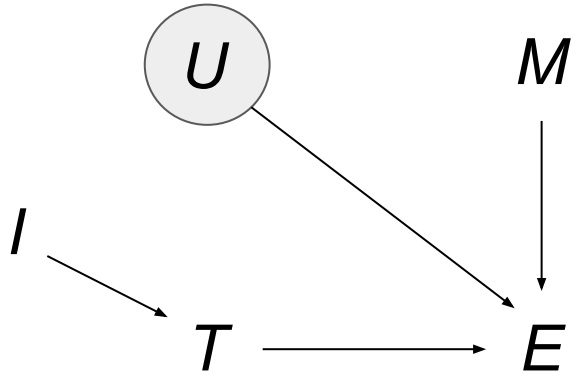
So is randomization the only way to render the ATE identified and construct unbiased estimates?



No Other Way?

So is randomization the only way to render the ATE identified and construct unbiased estimates?

No!

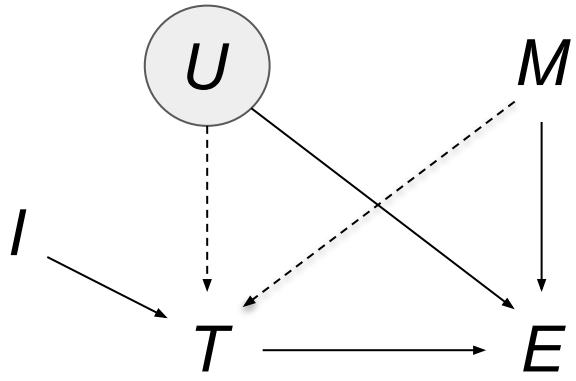


Instrumental Variables

I is an **instrumental variable** if (roughly)

- *I* is statistically independent of *U,M;*
- the only unblocked path from *I* to *E* goes through *T*

  (a path is blocked if it contains a sequence like … → *T* ← …).

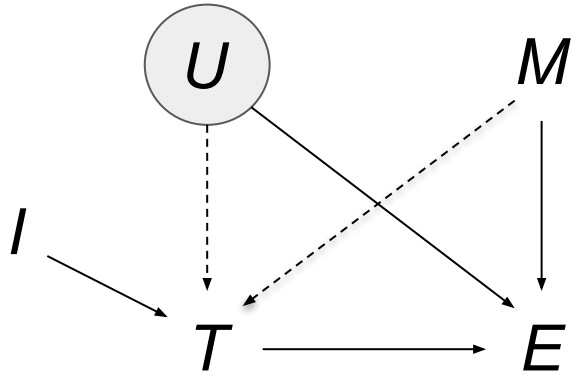


Instrumental Variables

Suppose that

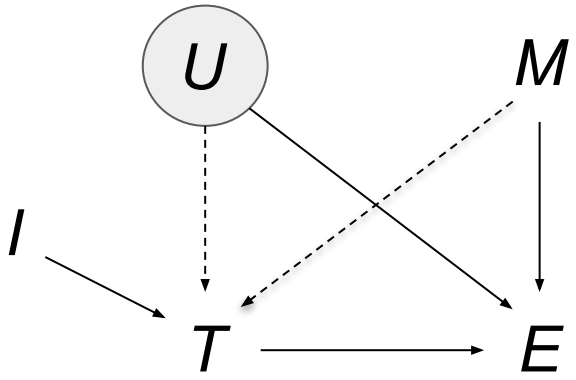
- physicians assign patients to treatment according to their therapeutic judgement
- and only consult a randomizing device (*I*) when they are in equipoise

then *I* is an instrumental variable.



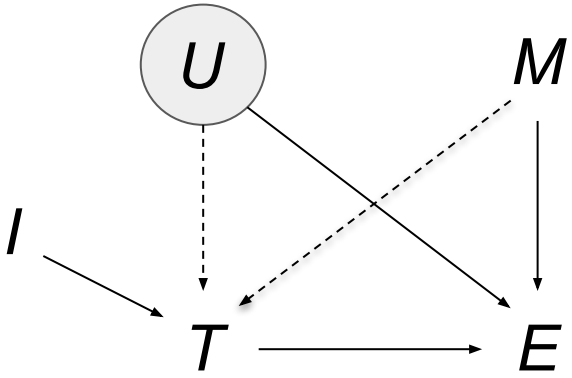
Instrumental Variables

Theorem (Angrist and Imbens 1995): When an instrumental variable satisfies a “monotonicity” condition, then the ATE is **identified** and there is an **unbiased estimator** of the ATE.



Instrumental Variables

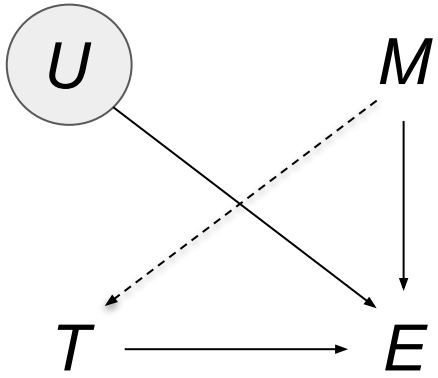
Theorem (Angrist and Imbens 1995): When an instrumental variable satisfies a “monotonicity” condition, then the ATE is **identified** and there is an **unbiased estimator** of the ATE.



Backdoor Adjustment

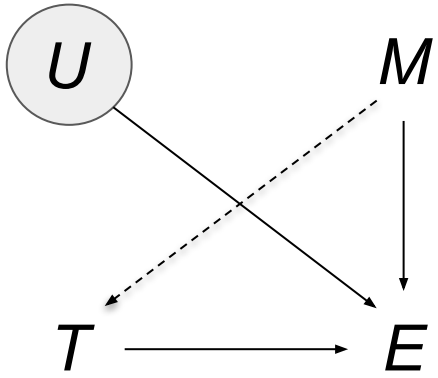
M satisfied the backdoor criterion w.r.t (*T, E)* if

- *M* is not a descendant of *T*;
- *M* blocks every path between *T* and *E* that has an arrow into *T.*

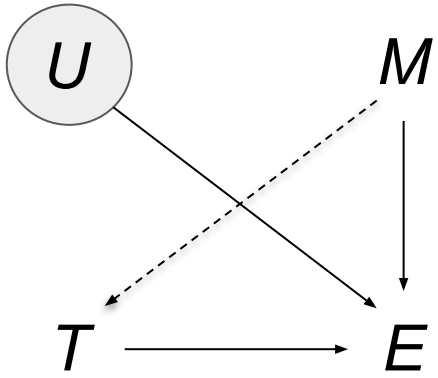


Backdoor Adjustment

Theorem (Pearl, 1993) If there is observed variable *Z* satisfying the backdoor criterion wrt (*T, E)*, then it is possible to construct an unbiased estimate of the causal effect of *T* on *E.*



Backdoor Adjustment

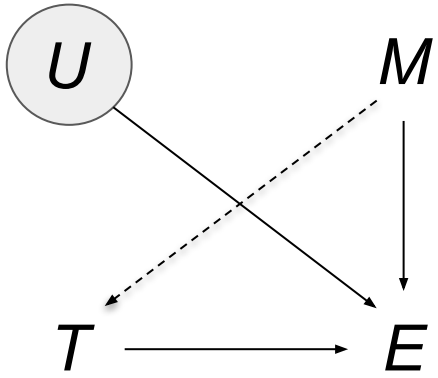


Suppose that

- physicians make assignment to treatment **only** on the basis of observed covariates *M*,

then *M* satisfies the backdoor criterion wrt (*T, E*).

Backdoor Adjustment



Suppose that

- physicians make assignment to treatment **only** on the basis of observed covariates *M*,

then *M* satisfies the backdoor criterion wrt (*T, E*).

Randomization On its Own Terms

Neither guaranteeing that

1. the ATE is identified, nor that
2. there exists an unbiased estimator of the ATE,

is sufficient to justify randomization.

Other designs get the same goods and are less hostile to individualized treatment.

Randomization On its Own Terms

The usual story establishes

1. the superiority of (quasi)**experimental** designs over **observational** designs;

but not

2. the superiority of **randomized** experimental designs over **other experimental designs.**

Randomization On its Own Terms

If there is a frequentist argument justifying randomization over other methods, it **cannot** be framed in terms of identifiability or unbiasedness of estimates.

It must be about **efficiency**.

I.e. the **variance** of the estimator.

Randomization On its Own Terms

Are there such arguments?

Randomization On its Own Terms

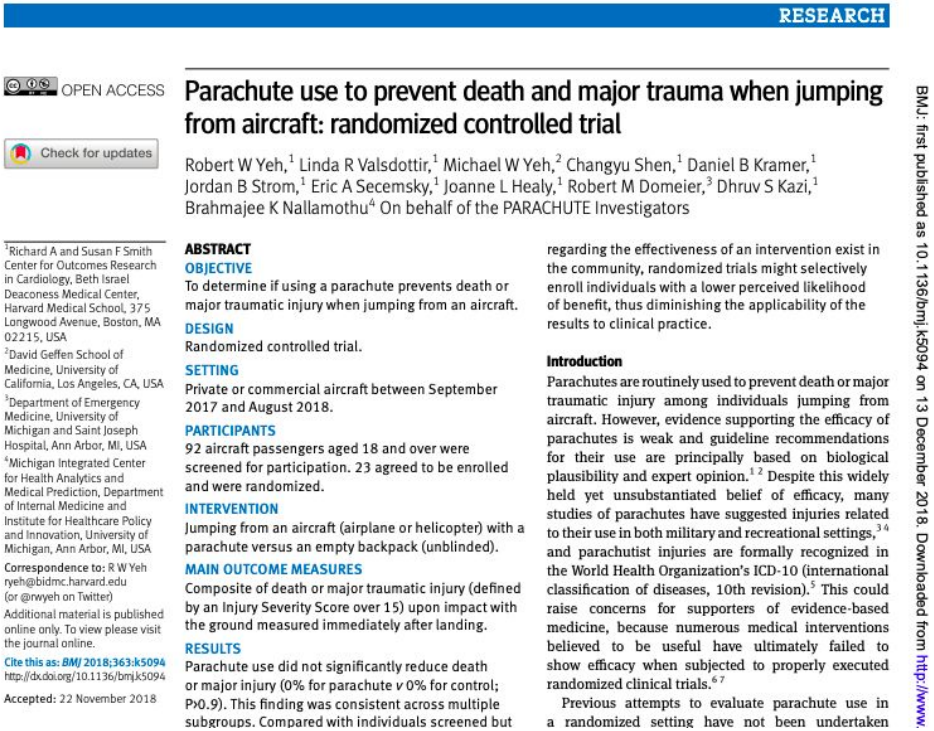
Are there such arguments?

There are definitely no **dominance** arguments: if you know that the disease is fatal without treatment, the variance-minimizing estimator of the ATE assigns everyone to treatment.

Randomization On its Own Terms

Are there such arguments?

There are definitely no **dominanc** fatal without treatment, the varian everyone to treatment.



RESEARCH

OPEN ACCESS

Check for updates

Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial

Robert W Yeh,[1] Linda R Valsdottir,[1] Michael W Yeh,[2] Changyu Shen,[1] Daniel B Kramer,[1] Jordan B Strom,[1] Eric A Secemsky,[1] Joanne L Healy,[1] Robert M Domeier,[3] Dhruv S Kazi,[1] Brahmajee K Nallamothu[4] On behalf of the PARACHUTE Investigators

[1]Richard A and Susan F Smith Center for Outcomes Research in Cardiology, Beth Israel Deaconess Medical Center, Harvard Medical School, 375 Longwood Avenue, Boston, MA 02215, USA
[2]David Geffen School of Medicine, University of California, Los Angeles, CA, USA
[3]Department of Emergency Medicine, University of Michigan and Saint Joseph Hospital, Ann Arbor, MI, USA
[4]Michigan Integrated Center for Health Analytics and Medical Prediction, Department of Internal Medicine and Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI, USA

Correspondence to: R W Yeh
ryeh@bidmc.harvard.edu
(or @rwyeh on Twitter)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2018;363:k5094
http://dx.doi.org/10.1136/bmj.k5094

Accepted: 22 November 2018

ABSTRACT

OBJECTIVE
To determine if using a parachute prevents death or major traumatic injury when jumping from an aircraft.

DESIGN
Randomized controlled trial.

SETTING
Private or commercial aircraft between September 2017 and August 2018.

PARTICIPANTS
92 aircraft passengers aged 18 and over were screened for participation. 23 agreed to be enrolled and were randomized.

INTERVENTION
Jumping from an aircraft (airplane or helicopter) with a parachute versus an empty backpack (unblinded).

MAIN OUTCOME MEASURES
Composite of death or major traumatic injury (defined by an Injury Severity Score over 15) upon impact with the ground measured immediately after landing.

RESULTS
Parachute use did not significantly reduce death or major injury (0% for parachute *v* 0% for control; P>0.9). This finding was consistent across multiple subgroups. Compared with individuals screened but regarding the effectiveness of an intervention exist in the community, randomized trials might selectively enroll individuals with a lower perceived likelihood of benefit, thus diminishing the applicability of the results to clinical practice.

Introduction

Parachutes are routinely used to prevent death or major traumatic injury among individuals jumping from aircraft. However, evidence supporting the efficacy of parachutes is weak and guideline recommendations for their use are principally based on biological plausibility and expert opinion.[1 2] Despite this widely held yet unsubstantiated belief of efficacy, many studies of parachutes have suggested injuries related to their use in both military and recreational settings,[3 4] and parachutist injuries are formally recognized in the World Health Organization's ICD-10 (international classification of diseases, 10th revision).[5] This could raise concerns for supporters of evidence-based medicine, because numerous medical interventions believed to be useful have ultimately failed to show efficacy when subjected to properly executed randomized clinical trials.[6 7]

Previous attempts to evaluate parachute use in a randomized setting have not been undertaken

Randomization On its Own Terms

Are there such arguments?

There are definitely no **dominance** arguments: if you know that the disease is fatal without treatment, the variance-minimizing estimator of the ATE assigns everyone to treatment.

There **might** be minimax arguments.

Optimal Design of Experiments

We have *N* experimental **units**

We have two **treatments** t_0 an

We write the **outcome** of unit *i*

We are interested in the **avera**

But we observe only exactly or

Kirstine Smith (1878-1939) was a Danish Statistician. She is credited with the creation of the field of optimal design of experiments. Karl Pearson considered her to be one of his most brilliant mathematical statisticians.

Her work with Pearson on minimum chi-square spurred a controversial dialog between Pearson and Fisher, and led to Fisher's introduction of sufficient statistics.

Selected Publications:

- Smith, K. (1916). On the 'best' values of the constants in frequency distributions. *Biometrika*, *11*(3), 262–276.
- Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, *12*(1/2), 1–85.
- Smith, K. (1922). The standard deviations of fraternal and parental correlation coefficients. *Biometrika*, *14*(1/2), 1–22.

Optimal Design of Experiments

We have *N* experimental **units**, e.g. plots of land, or patients in a trial.

We have two **treatments** t_0 and t_1, e.g. varieties of wheat, competing drugs.

We write the **outcome** of unit *i* under treatment *t* as y_i^t.

We are interested in the **average treatment effect** $\alpha := N^{-1} \Sigma_i E[y_i^1] - E[y_i^0]$.

But we observe only exactly one of $\{ y_i^1 , y_i^0 \}$.

Optimal Design of Experiments

A **design** is an assignment of units to treatments, i.e. a function $f : N \rightarrow \{0, 1\}$.

Let D be the set of all designs.

Let $y_f = (y_1, \ldots, y_N)$ be the observations arising from the design f.

We have an unbiased **estimator** $\hat{a}(y_f)$, usually the difference-of-means.

The **loss** is a random variable $L_f = L(\alpha, \hat{a}(y_f))$.

ODE: pick the design f in D that minimizes $E[L_f]$.

Optimal Design of Experiments

The theory of optimal design considers only **deterministic** designs!

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 539, 1452–1465: Theory and Methods
https://doi.org/10.1080/01621459.2020.1863221

Minimax Efficient Random Experimental Design Strategies With Application to Model-Robust Design for Prediction

Timothy W. Waite[a] and David C. Woods[b]

[a]Department of Mathematics, University of Manchester, Manchester, UK; [b]Statistical Sciences Research Institute, University of Southampton, Southampton, UK

Optimal Design of Experiments

The theory of optimal design considers only **deterministic** designs!

Though the above property appears to give a strong argument in favor of the use of ξ_{mM}, in fact in both game theory and statistical decision theory it is widely recognized that a minimax deterministic decision is often outperformed by a randomized decision strategy (e.g., Blackwell and Girshick 1979, Berger 1985, chap. 5, Thie and Keough 2011, chap. 9). Since design selection can be viewed as a decision problem, or alternatively a game pitting the Statistician against Nature, it stands to reason that random decision strategies should also be beneficial for experimental design. Nonetheless, aside from a few minimax analyses of Fisherian randomization (Wu 1981; Li 1983; Hooper 1989; Bhaumik and Mathew 1995), the topic of minimax random strategies for design selection appears almost totally unexplored in the literature.

Optimal Design of Experiments

The theory of optimal design considers only **deterministic** designs!

Though the above property appears to give a strong argument in favor of the use of ξ_{mM}, in fact in both game theory and statistical decision theory it is widely recognized that a minimax deterministic decision is often outperformed by a randomized decision strategy (e.g., Blackwell and Girshick 1979, Berger 1985, chap. 5, Thie and Keough 2011, chap. 9). Since design selection can be viewed as a decision problem, or alternatively a game pitting the Statistician against Nature, it stands to reason that random decision strategies should also be beneficial for experimental design. Nonetheless, aside from a few minimax analyses of Fisherian randomization (Wu 1981; Li 1983; Hooper 1989; Bhaumik and Mathew 1995), the topic of minimax random strategies for design selection appears almost totally unexplored in the literature.

Minimax Justifications

A series of somewhat neglected papers (Wu 1981; Li 1983; Waite and Woods 2020; Bai 2021) develops a **minimax risk** argument for randomization.

Minimax Justifications Redux: Causal States

A **causal state** is a random *N* x 2 matrix **Y** of potential outcomes, where

$$\mathbf{Y}_i = (Y_i(0), Y_i(1))$$

represents the counterfactual outcome under control and treatment, respectively, for the i^{th} patient.

Let $\mathcal{Y}$ be the set of causal states that the researchers consider a priori possible.

Minimax Justifications Redux: Permutation Group

Let Π be a collection of permutations of *N.*

Assumption 1: if **Y** is in $\mathcal{Y}$ and π is in Π, then π**Y** is in $\mathcal{Y}$. (**Closure under** Π)

The permutation group partitions people into “clinically equivalent” strata, e.g.

{<45 and severely ill, <45 and mildly ill, >=45 and severely ill, >=45 and mildly ill}

Minimax Justifications Redux: Strategies

A **deterministic** design is a *k x n* binary matrix *T* in which a row specifies which subjects receive treatment.

A deterministic design *T* and a state **Y** determine a random observed outcome

$$\mathbf{Y}_T = diag(\mathbf{Y} \cdot T).$$

An **estimator** (for *θ*) is a function $\hat{\theta}(T, y) \in \mathbb{R}^d$.

Minimax Justifications Redux: Strategies

A **strategy** is a pair $(T, \hat{\theta})$ of a design and an estimator. Let $\mathscr{S}$ be the set of all feasible strategies.

Assumption 2: If $(T, \hat{\theta}) \in \mathcal{S}$, then $(\pi T, \hat{\theta}) \in \mathcal{S}$.

If we can treat subject *i,* then we can also treat subjects indistinguishable from *i.*

Assumption 3: $\hat{\theta}(T, y) = \hat{\theta}(\pi T, \pi y)$

Renaming equivalent patients doesn't change the value of the estimate.

Minimax Justifications Redux: Loss Functions

A **loss function** is a function $L : \mathcal{Y} \times \mathbb{R}^d \mapsto \mathbb{R}^{\geq 0}$.

Assumption 4: $L(\mathbf{Y}, \mathbf{x}) = \mathbf{L}(\pi\mathbf{Y}, \mathbf{x})$.

For example, since **Y** and π**Y** agree in the value of the ATE, this assumption is satisfied by the usual strategies.

Note: If your loss function doesn't depend on **Y** (for example: clinical loss), then this assumption is also satisfied.

Minimax Justifications Redux: Expected Loss

If **Y** is a state, **T** is a (randomized) design and $\hat{\theta}$ is an estimator, then the **risk** (expected loss) is:

$$r_{\mathbf{Y}}(\hat{\theta}, \mathbf{T}) := E[L(\mathbf{Y}, \hat{\theta}(\mathbf{T}, \mathbf{Y_T}))].$$

A design **T** is **ancillary** if it is independent from **Y.**

A First Minimax Theorem

Let **Π** be a random permutation taking values in Π.

Theorem 1. Suppose that Assumptions 1-4 hold. Suppose **T** is ancillary and **Π** is independent of (**Y, T**). Then:

$$\sup_{\mathbf{Y} \in \mathcal{Y}} r_{\mathbf{Y}}(\hat{\theta}, \mathbf{\Pi T}) \leq \sup_{\mathbf{Y} \in \mathcal{Y}} r_{\mathbf{Y}}(\hat{\theta}, \mathbf{T}).$$

Corollary: in the worst case, a randomized design is at least as good as any deterministic design.

More Minimax Theory

The preceding holds very generally. Can we say something more specific for the case when θ is the ATE and L is the usual squared-error loss?

More Minimax Theory

For deterministic potential outcomes and fully randomized designs, Imbens and Rubin (2015) prove that the loss of the usual difference-of-means is given by:

$$S_t^2\left(\frac{1}{n_t}-\frac{1}{N}\right)+S_c^2\left(\frac{1}{n_c}-\frac{1}{N}\right)+\frac{2}{N(N-1)}\sum_{i=1}^{N}(Y_i(0)-\overline{Y}(0))(Y_i(1)-\overline{Y}(1).$$

Where:

$$S_t^2=\frac{1}{N-1}\cdot\mathbb{E}\left[\sum_{i=1}^{N}(Y_i(t)-\overline{Y}(t))^2\right]$$

$$S_c^2=\frac{1}{N-1}\cdot\mathbb{E}\left[\sum_{i=1}^{N}(Y_i(c)-\overline{Y}(c))^2\right]$$

More Minimax Theory

Suppose that all patients are equivalent and that for all *i*

$$|Y_i(0) - \overline{Y}(0)| \in [a, b] \text{ and } |Y_i(1) - \overline{Y}(1)| \in [c, d],$$

then the minimax fully randomized design adjusts treatment number to the maximum values of S_t^2 and S_c^2.

If all patients are not equivalent, the same idea applies to blocked designs: adjust block treatment numbers to worst-case treatment variability in each block.

More Minimax Theory

What about **balance**?

Suppose that $\mathcal{Y}$ is closed under **column** permutation.

Then the minimax (block) randomized design is given by balanced (block) randomization.

Ongoing Work

Can we generalize away from the standard estimators? In other words: can we simultaneously optimize for the worst case design-estimator combo?

The answer: yes! But only if we restrict ourselves to **unbiased** estimators. (Then we can even drop Assumption 2.)