What works in Boston may not work in Los Angeles: Understanding site differences and generalizing effects from one site to another.

Kara Rudolph
with Mark van der Laan

RWJF Health and Society Scholar
UC Berkeley / UC San Francisco
Outline

1. Motivation
   - Motivating example

2. Methodologic Challenges

3. Approach

4. Results

5. Future directions

Kara Rudolph (UCB/UCSF)
Motivation

Should we expect that a policy/program/intervention implemented in one place will have the same effect when implemented in another place?
Motivation

- Should we expect that a policy/program/intervention implemented in one place will have the same effect when implemented in another place?
- Not always.
Motivation

- Should we expect that a policy/program/intervention implemented in one place will have the same effect when implemented in another place?
- Not always.
  - Differences in site-level variables (e.g., implementation, economy) that modify intervention effectiveness, AND/OR
Motivation

Should we expect that a policy/program/intervention implemented in one place will have the same effect when implemented in another place?

Not always.

1. Differences in site-level variables (e.g., implementation, economy) that modify intervention effectiveness, AND/OR

2. Differences in person-level variables (i.e., population composition) that modify intervention effectiveness.
Motivation

- Budgets are limited.
Motivation

- Budgets are limited.
- Need to target the policy/intervention to those places that stand to benefit most.

Research question:
What do we expect the effect of an intervention to be in a new place, accounting for population composition?
Motivation

- Budgets are limited.
- Need to target the policy/intervention to those places that stand to benefit most.
  - E.g., planned expansion of intervention. Where should it be expanded to have the largest effect? How is success defined?
Motivation

- Budgets are limited.
- Need to target the policy/intervention to those places that stand to benefit most.
  - E.g., planned expansion of intervention. Where should it be expanded to have the largest effect? How is success defined?
  - Can you think of any practical examples of this?

Research question:
What do we expect the effect of an intervention to be in a new place, accounting for population composition?
Motivation

- Budgets are limited.
- Need to target the policy/intervention to those places that stand to benefit most.
  - E.g., planned expansion of intervention. Where should it be expanded to have the largest effect? How is success defined?
  - Can you think of any practical examples of this?
- Research question: **What do we expect the effect of an intervention to be in a new place, accounting for population composition?**
Outline

1. Motivation
   - Motivating example

2. Methodologic Challenges

3. Approach

4. Results

5. Future directions
Motivating Example

- Moving To Opportunity (MTO)\textsuperscript{1}

In discussing differences in effects across sites, MTO researchers concluded:

> Of course, if it had been possible to attribute differences in impacts across sites to differences in site characteristics, that would have been very valuable information. Unfortunately, that was not possible. With only five sites, which differ in innumerable potentially relevant ways, it was simply not possible to disentangle the underlying factors that cause impacts to vary across sites. (This is true for both the quantitative analysis and for any qualitative analysis of the impacts that might be undertaken.)

Why are the researchers saying this? Do you agree?

---

Motivating Example

- **Research Question (MTO-specific):** Are differences in intervention effects across cities due to differences in implementation? City-level differences (e.g., the economy)? Or differences in population composition?
Outline

1 Motivation
   • Motivating example

2 Methodologic Challenges

3 Approach

4 Results

5 Future directions
Typically, multi-site data are analyzed using fixed effects.
Typically, multi-site data are analyzed using fixed effects.

- Usually assumes that we answered “Yes” to whether we expect the intervention effect in one site is the same as the other site.
Typically, multi-site data are analyzed using fixed effects.

- Usually assumes that we answered “Yes” to whether we expect the intervention effect in one site is the same as the other site.

- Why is that the case?
Typically, multi-site data are analyzed using fixed effects.

- Usually assumes that we answered “Yes” to whether we expect the intervention effect in one site is the same as the other site.
- Why is that the case?
- Dummy variables for site changes the intercept but not the treatment effect coefficient. Assume that the conditional effect (regression coefficient) of the intervention in one site is the same as in another site.
Typically, multi-site data are analyzed using fixed effects.

- Usually assumes that we answered “Yes” to whether we expect the intervention effect in one site is the same as the other site.
- Why is that the case?
- Dummy variables for site changes the intercept but not the treatment effect coefficient. Assume that the conditional effect (regression coefficient) of the intervention in one site is the same as in another site.

Need to apply the results from one city/site to a target city/site based on the observed population composition.
Typically, multi-site data are analyzed using fixed effects.

- Usually assumes that we answered “Yes” to whether we expect the intervention effect in one site is the same as the other site.

- Why is that the case?

- Dummy variables for site changes the intercept but not the treatment effect coefficient. Assume that the conditional effect (regression coefficient) of the intervention in one site is the same as in another site.

Need to apply the results from one city/site to a target city/site based on the observed population composition.

- Transportability/ generalizability/ external validity.
Most common: Use fixed effects for site.
What’s been done

- Most common: Use fixed effects for site.
  - Conditional effect is not as policy relevant as marginal effect

---

What’s been done

- Most common: Use fixed effects for site.
  - Conditional effect is not as policy relevant as marginal effect
  - We usually don’t believe the assumption

---

What’s been done

- Most common: Use fixed effects for site.
  - Conditional effect is not as policy relevant as marginal effect
  - We usually don’t believe the assumption

- Common-ish: Post-stratification/direct standardization. E.g., age-adjusted rates of disease for comparisons between populations.

---

What’s been done

- Most common: Use fixed effects for site.
  - Conditional effect is not as policy relevant as marginal effect
  - We usually don’t believe the assumption

- Common-ish: Post-stratification/ direct standardization. E.g., age-adjusted rates of disease for comparisons between populations.
  - Breaks down with continuous characteristics or multiple characteristics because of small cell sizes

---

What's been done

- Most common: Use fixed effects for site.
  - Conditional effect is not as policy relevant as marginal effect
  - We usually don’t believe the assumption

- Common-ish: Post-stratification/ direct standardization.\(^3\) E.g., age-adjusted rates of disease for comparisons between populations.
  - Breaks down with continuous characteristics or multiple characteristics because of small cell sizes
  - No standard errors/ no inference

What’s been done

- Less common, rare: Model-based approaches: Horvitz-Thompson weighting (model-based standardization),\(^4\) propensity score matching,\(^5\) and principal stratification\(^6\)

---


What’s been done

  - Relies on correct model specification

---


What’s been done

- Less common, rare: Model-based approaches: Horvitz-Thompson weighting (model-based standardization),\textsuperscript{4} propensity score matching,\textsuperscript{5} and principal stratification\textsuperscript{6}
  
  - Relies on correct model specification
  - Inference with machine learning is unclear


\textsuperscript{7} Pearl, J. & Bareinboim, E. \textit{Transportability across studies: A formal approach} tech. rep. (DTSICDocument 2011).
What’s been done

- Less common, rare: Model-based approaches: Horvitz-Thompson weighting (model-based standardization), propensity score matching, and principal stratification
  - Relies on correct model specification
  - Inference with machine learning is unclear
  - With exception of principal stratification, have not been extended to encouragement-design interventions

---


What’s been done

  - Relies on correct model specification
  - Inference with machine learning is unclear
  - With exception of principal stratification, have not been extended to encouragement-design interventions

- Pearl and Bareinbom: formalized theory and assumptions for transportability.

---


Outline

1. Motivation
   - Motivating example

2. Methodologic Challenges

3. Approach

4. Results

5. Future directions
Our contribution

New statistical method for “transporting” effects from one population to another

- Transport formula for multi-site encouragement-design interventions (extending Pearl and Bareinboim’s work).

---

Our contribution

New statistical method for “transporting” effects from one population to another\(^8\)

- Transport formula for multi-site encouragement-design interventions (extending Pearl and Bareinboim’s work).
- Estimation using transport formulas addressing previous gaps:

---

Our contribution

New statistical method for “transporting” effects from one population to another\(^8\)

- Transport formula for multi-site encouragement-design interventions (extending Pearl and Bareinboim’s work).
- Estimation using transport formulas addressing previous gaps:
  - Inference based on theory (even when using machine learning)

Our contribution

New statistical method for “transporting” effects from one population to another

- Transport formula for multi-site encouragement-design interventions (extending Pearl and Bareinboim’s work).
- Estimation using transport formulas addressing previous gaps:
  - Inference based on theory (even when using machine learning)
  - Double robust: can misspecify multiple models and still get unbiased estimates

---

Problem: there are a lot of relationships to specify and we don’t know the truth!

Can you guess the correct models when $W$ is high dimensional? All interactions? Correct form (e.g., linear, quadratic, spline)?

Note: $A =$ instrument/encouragement, $Z =$ exposure, $Y =$ outcome, $S =$ site, $W =$ covariates/characteristics
Targeted maximum likelihood estimators (TMLE) for the following estimands:

- Effect of A on Y (intent-to-treat)
- Effect of Z on Y using randomization of the instrument (complier average treatment effect)
- Effect of Z on Y ignoring randomization

Note: A = instrument/encouragement, Z = exposure, Y = outcome, S = site, W = covariates/characteristics
In everyday language, what does TMLE do?

1. Start with identifying the parameter you’re interested in estimating. E.g., the ITTATE, $\psi$.

2. Get initial estimate for $\psi$. E.g., run a regression of the $Y$ model setting $A = 1$ and $A = 0$. The difference will be the initial estimate.

3. The $Y$ model may not be perfect. (If it is, you’re done.) This initial estimate is then adjusted by something called the clever covariate, $C$, which is derived from the efficient influence curve. It uses information from the other models improve upon the initial estimate.

4. This fluctuation can be iterated until convergence.
In everyday language, what does TMLE do?

Initial estimate $\rightarrow + C_1 \rightarrow + C_2 \rightarrow$ True estimate $\rightarrow$ TMLE estimate
Outline

1 Motivation
   - Motivating example

2 Methodologic Challenges

3 Approach

4 Results

5 Future directions
Performance

- Results for intent-to-treat effect of A on Y. Results are similar for the two other estimators.

<table>
<thead>
<tr>
<th>Model specification</th>
<th>% Bias</th>
<th>Variance</th>
<th>Coverage</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>All models correct</td>
<td>-0.67</td>
<td>0.0004</td>
<td>95.01</td>
<td>0.0004</td>
</tr>
<tr>
<td>S model misspecified</td>
<td>-0.49</td>
<td>0.0004</td>
<td>95.34</td>
<td>0.0004</td>
</tr>
<tr>
<td>Z model misspecified</td>
<td>-0.67</td>
<td>0.0004</td>
<td>95.00</td>
<td>0.0004</td>
</tr>
<tr>
<td>Y model misspecified</td>
<td>-0.71</td>
<td>0.0005</td>
<td>95.36</td>
<td>0.0005</td>
</tr>
<tr>
<td>S,Z models misspecified</td>
<td>-0.49</td>
<td>0.0004</td>
<td>95.29</td>
<td>0.0004</td>
</tr>
<tr>
<td>S,Z,Y models misspecified</td>
<td>6.05</td>
<td>0.0004</td>
<td>94.84</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

Note: A = instrument/encouragement, Z = exposure, Y = outcome, S = site, W = covariates/characteristics
Sensitivity to positivity violations

- Structural positivity violations: Person with some set of covariate values in one treatment/selection group has a zero probability of being in another treatment/selection group.
Sensitivity to positivity violations

- **Structural positivity violations:** Person with some set of covariate values in one treatment/selection group has a zero probability of being in another treatment/selection group.

- **Practical positivity violations:** This probability isn’t strictly zero, but it’s close.
Sensitivity to positivity violations

- **Structural positivity violations**: Person with some set of covariate values in one treatment/selection group has a zero probability of being in another treatment/selection group.

- **Practical positivity violations**: This probability isn’t strictly zero, but it’s close.

- Why is this a problem?
Sensitivity to positivity violations

- Practical positivity violations are a substantial issue in real world data.
Sensitivity to positivity violations

- Practical positivity violations are a substantial issue in real world data.
- Why might we expect it in the example below?

![Graph showing predicted probability of job strain](image-url)
Sensitivity to positivity violations

Which of the 3 estimands is most vulnerable to these violations using the MTO data?
Sensitivity to positivity violations

- Which of the 3 estimands is most vulnerable to these violations using the MTO data?
- What are some other real-world examples that might be vulnerable to positivity violations?
## Sensitivity to positivity violations

<table>
<thead>
<tr>
<th>EATE</th>
<th>$C_Y(A = 1)$</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data-generating mechanism 1</td>
<td>$0.49(0.38)$</td>
<td>$0.05$</td>
<td>$2.46$</td>
</tr>
<tr>
<td>Data-generating mechanism 2</td>
<td>$1.07(1.62)$</td>
<td>$0.15 \times 10^{-2}$</td>
<td>$26.26$</td>
</tr>
<tr>
<td>Application</td>
<td>$2.05(2.76)$</td>
<td>$4.54 \times 10^{-2}$</td>
<td>$13.11$</td>
</tr>
</tbody>
</table>
Sensitivity to positivity violations

<table>
<thead>
<tr>
<th>Specification</th>
<th>%Bias</th>
<th>SE $\sqrt{n}$ (1.60)</th>
<th>Cov</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EATE: Without Positivity Violations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All models correct</td>
<td>-0.31</td>
<td>1.60</td>
<td>94.94</td>
<td>0.0005</td>
</tr>
<tr>
<td>S model misspecified</td>
<td>-0.38</td>
<td>1.46</td>
<td>93.68</td>
<td>0.0005</td>
</tr>
<tr>
<td>Z model misspecified</td>
<td>-0.31</td>
<td>1.48</td>
<td>93.01</td>
<td>0.0005</td>
</tr>
<tr>
<td>Y model misspecified</td>
<td>-0.29</td>
<td>1.62</td>
<td>95.09</td>
<td>0.0005</td>
</tr>
<tr>
<td>S,Z models misspecified</td>
<td>-0.43</td>
<td>1.36</td>
<td>92.95</td>
<td>0.0004</td>
</tr>
<tr>
<td>S,Z,Y models misspecified</td>
<td>14.46</td>
<td>1.37</td>
<td>76.27</td>
<td>0.0009</td>
</tr>
<tr>
<td><strong>EATE: With Positivity Violations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All models correct</td>
<td>0.18</td>
<td>3.60</td>
<td>91.36</td>
<td>0.0029</td>
</tr>
<tr>
<td>S model misspecified</td>
<td>1.98</td>
<td>1.96</td>
<td>86.33</td>
<td>0.0012</td>
</tr>
<tr>
<td>Z model misspecified</td>
<td>0.18</td>
<td>2.67</td>
<td>82.93</td>
<td>0.0029</td>
</tr>
<tr>
<td>Y model misspecified</td>
<td>2.09</td>
<td>4.17</td>
<td>96.05</td>
<td>0.0027</td>
</tr>
<tr>
<td>S,Z models misspecified</td>
<td>2.18</td>
<td>1.38</td>
<td>79.27</td>
<td>0.0009</td>
</tr>
<tr>
<td>S,Z,Y models misspecified</td>
<td>-52.11</td>
<td>1.41</td>
<td>2.49</td>
<td>0.0065</td>
</tr>
</tbody>
</table>
Strategies for addressing positivity violations

- Limit the sample to the area of support
Strategies for addressing positivity violations

- Limit the sample to the area of support
- Truncate weights
Strategies for addressing positivity violations

- Limit the sample to the area of support
- Truncate weights
- Exclude covariates that are neither 1) confounders of the exposure-outcome relationship, nor 2) affect transportability.
Strategies for addressing positivity violations

- Limit the sample to the area of support
- Truncate weights
- Exclude covariates that are neither 1) confounders of the exposure-outcome relationship, nor 2) affect transportability.
- Moving the weights from the clever covariate into the model fitting step
## Strategies for addressing positivity violations

<table>
<thead>
<tr>
<th>Truncation Level</th>
<th>%Bias</th>
<th>SE\times\sqrt{n}</th>
<th>Cov</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EATE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No modification</td>
<td>0.18</td>
<td>3.60</td>
<td>91.36</td>
<td>0.0029</td>
</tr>
<tr>
<td>Truncation at 0.01/100</td>
<td>2.29</td>
<td>3.23</td>
<td>92.50</td>
<td>0.0024</td>
</tr>
<tr>
<td>Truncation at 0.05/20</td>
<td>2.71</td>
<td>2.40</td>
<td>89.78</td>
<td>0.0016</td>
</tr>
<tr>
<td>Truncation at 0.1/10</td>
<td>2.60</td>
<td>1.90</td>
<td>84.96</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

Kara Rudolph (UCB/UCSF)
Can our new statistical method shed light on the previously intractable problem of not knowing why there are differences in effects across sites?

We take two of the sites: LA and Boston. Outcome: adolescent school drop out at follow-up. We use full data from Boston. We ignore the outcome data from LA. Using the outcome model from Boston, we predict the intervention effect in LA, accounting for differences in population composition between the two cities.
Can our new statistical method shed light on the previously intractable problem of not knowing why there are differences in effects across sites?

We take two of the sites: LA and Boston.
Can our new statistical method shed light on the previously intractable problem of not knowing why there are differences in effects across sites?

We take two of the sites: LA and Boston.

Outcome: adolescent school drop out at follow-up.
Can our new statistical method shed light on the previously intractable problem of not knowing why there are differences in effects across sites?

We take two of the sites: LA and Boston.

Outcome: adolescent school drop out at follow-up.

We use full data from Boston. We ignore the outcome data from LA. Using the outcome model from Boston, we predict the intervention effect in LA, accounting for differences in population composition between the two cities.
Results

Real results: Boston

Intervention Effect on Risk of School Drop Out

Kara Rudolph (UCB/UCSF)

Generalizing effects across sites
Results

Predicted results: LA

![Graph showing intervention effect on risk of school drop out for different sites and treatment methods.]
Predicted vs. real results: LA

![Graph showing intervention effect on risk of school drop out for Boston, LA, and transported LA, TMLE across ITTATE and CATE categories.](image-url)
Results

- The transported estimates for LA are similar to true LA estimates.
The transported estimates for LA are similar to true LA estimates.

Using population composition, we can predict the effect for LA → intervention effect on school dropout is transportable.
Results

- The transported estimates for LA are similar to true LA estimates.
- Using population composition, we can predict the effect for LA \(\rightarrow\) intervention effect on school dropout is transportable.
- This means that the difference in effects between Boston and LA can be largely explained by population composition.
Aside: the importance of incorporating machine learning

- Difference of probability of staying in school

- Model:
  - none
  - parametric
  - superlearner

- Sites:
  - Boston
  - LA
  - Transported LA, TMLE

- Kara Rudolph (UCB/UCSF)

Generalizing effects across sites
Superlearner\(^9\)

- Ensemble machine learning
- Weights multiple machine learning algorithms to get best prediction
- Guaranteed to perform at least as well as best algorithm included in the weighting

We should not expect an intervention/program/policy to have the same effect in one city as in another city.
We should not expect an intervention/program/policy to have the same effect in one city as in another city.

In an era of shrinking budgets, important to recognize that what works in Boston may not work in LA, so resources can be targeted optimally.
We should not expect an intervention/program/policy to have the same effect in one city as in another city.

In an era of shrinking budgets, important to recognize that what works in Boston may not work in LA, so resources can be targeted optimally.

Broadly useful: multi-site epidemiologic studies, large-scale policy or program interventions, clinical trials.
Future Directions

- Examine other strategies to reduce sensitivity to practical positivity violations, especially excluding covariates and moving the weights.
Future Directions

- Examine other strategies to reduce sensitivity to practical positivity violations, especially excluding covariates and moving the weights.
- In-depth application of transportability to MTO to understand the relationship between neighborhood poverty and exposure to violence and violent behaviors.
Future Directions

- Examine other strategies to reduce sensitivity to practical positivity violations, especially excluding covariates and moving the weights.
- In-depth application of transportability to MTO to understand the relationship between neighborhood poverty and exposure to violence and violent behaviors.
- Grant application to extend the transportability method to mediation mechanisms. Examine mediation of the relationship between neighborhood poverty on adolescent risk behaviors by the school environment.
Future Directions

- Examine other strategies to reduce sensitivity to practical positivity violations, especially excluding covariates and moving the weights.

- In-depth application of transportability to MTO to understand the relationship between neighborhood poverty and exposure to violence and violent behaviors.

- Grant application to extend the transportability method to mediation mechanisms. Examine mediation of the relationship between neighborhood poverty on adolescent risk behaviors by the school environment.

- Other ideas? Suggestions?
How can I do this?

- Use the R functions that I wrote
  - Parametric or semiparametric options

\[
\text{ittatetmle} \leftarrow \text{function}\left(a, z, y, \text{site}, w, \text{truncate}, l\text{bound}\right)
\]

\[
\text{catetmle} \leftarrow \text{function}\left(ca, cz, cy, c\text{site}, cw, c\text{truncate}, c\text{lbou}nd\right)
\]

\[
\text{noinstratetmle} \leftarrow \text{function}\left(a, z, y, \text{site}, w, \text{truncate}, l\text{bound}\right)
\]
Thanks!

www.biostat.jhsph.edu/~krudolph
kara.rudolph@berkeley.edu

Robert Wood Johnson Foundation Health & Society Scholars program, UCSF/UCB

Collaborators

- Mark van der Laan, UC Berkeley
- Jennifer Ahern, UC Berkeley,
- Maria Glymour, UCSF
- Theresa Osypuk, University of Minnesota