Causal inference with skewed outcome data: Moving beyond the "ignore or transform" approach

Daisy A. Shepherd[*1,2] **& Margarita Moreno-Betancur**[1,2]

[1]Clinical Epidemiology & Biostatistics Unit, Murdoch Children's Research Institute, [2]Department of Paediatrics, The University of Melbourne

[*]daisy.shepherd@mcri.edu.au

Background & Motivation

- Causal effects (CE) are typically defined as a contrast of mean potential outcomes under exposure versus no exposure.
- There are many established methods to adjust for potential confounding bias due to the lack of randomisation.
- However, epidemiological studies may suffer from skewed continuous outcome data, for which the mean may no longer be a meaningful summary statistic.
- **But how do we estimate causal effects for skewed outcomes?**

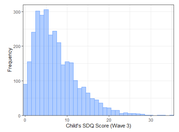

Figure 1. Distribution of outcome for LSAC example [1].

1. **IGNORE** the skewness and use the mean as summary statistic?
   - ✓ Appropriate when the mean remains of interest despite the skewness
   - ✓ Allows established confounding-adjustment methods to be applied
   - × Not appropriate if interested in central value of the outcome distribution
2. **TRANSFORM** the outcome to a more symmetric distribution?
   - ✓ Mean is interpretable and can apply established methods
   - × A suitable transformation may not exist
   - × More complex interpretation of CEs than on the original scale
3. **DEFINE** the causal effect as a contrast of median potential outcomes?
   - ✓ Widely acknowledged definition of CE [2]
   - × Limited understanding of confounding-adjustment methods
   - × Application of methods in practice is scarce

Study Aim: Describe and evaluate confounding-adjustment methods to estimate the causal difference in median potential outcomes, with the aim to increase understanding of their utility and encourage application in practice where appropriate.

Proposed Methods

- We consider an observational study with continuous skewed outcome variable Y, a binary exposure variable X, and a vector of confounder variables $\mathbf{C}$.
- The causal effect of interest δ is defined as the difference between the median counterfactual outcome under each exposure level: $\delta = m[Y^{x=1}] - m[Y^{x=0}]$
- To estimate δ from observational data, we identified three singly-robust confounding-adjustment methods in the literature (methods 1-3), alongside a proposed method (4).

Method 1: Multivariable Quantile Regression

1. Fit a quantile regression (QR) model of Y conditional on X and $\mathbf{C}$.
2. Using the 50th quantile, the coefficient for X is an unbiased estimate of δ (under certain assumptions)[3].

Method 2: Weighted Quantile Regression

→ Uses the framework of inverse-probability weighting to create a pseudo-population.

1. Fit a QR model $Y \sim X$, with observations weighted inversely proportional to the propensity score.
2. Using the 50th quantile, the coefficient for X is a consistent estimator of δ (under certain assumptions)[3].

Method 3: IPW Estimator

→ Like method 2, uses the framework of inverse-probability weighting.
→ Instead of using a weighted QR model, uses a direct derivation of a weighted estimator[4], by solving for y under X=0,1 weighting observations by $W_{x,i}$.

$$\sum_{i=1}^{n} W_{x,i} I(Y_i \leq y) = 0.5$$

Method 4: Approximate G-computation

→ Heuristically motivated based on G-computation used when the CE is defined using the mean.

1. Fit a QR model of Y conditional on X and $\mathbf{C}$.
2. Predict the median outcome under X=0,1 for every observation.
3. Aggregate predictions across each exposure group to estimate $m[Y^{x=1}]$ and $m[Y^{x=0}]$.

→ We conjecture δ is approximated as the difference in aggregated predictions under each exposure level.
→ We considered two aggregation approaches:
G-comp (mean) = calculate the mean of predictions
G-comp (med) = calculate the median of predictions

Simulation Study

Objective: Investigate the performance of each confounding-adjustment method in a realistic setting and under varying degrees of skewness in the outcome variable.

1. Study Design

- Motivated by an observational study of kindergarten children (LSAC[4]) estimating the causal effect of maternal mental health (binary indicator) on a child's behaviour in early childhood (measured by SDQ score; positively skewed, Figure 1).
- Generated datasets consisting of:
  - X: Exposure variable (binary)
  - Y: Skewed outcome variable (continuous), drawn from a log-normal distribution
  - **C**: Confounder variables (3 binary, 2 continuous)

Explored 4 skewness scenarios (SS)
→ Different degrees of skewness in Y
→ SS1 = weakest, SS4=strongest

→

Per SS:
1000 datasets of
1000 observations

→

Replicated under
weak and strong
confounding

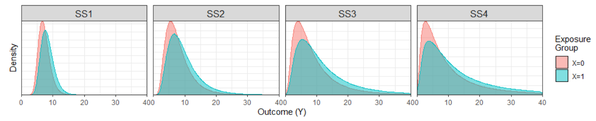

Figure 2. Log-normal distributions used to generate Y under each skewness scenario.

2. Simulation Results

- Results were similar for both levels of confounding, so weak (relative bias fixed at 10%) presented only.
- Standard errors and confidence intervals were estimating using a non-parametric bootstrap approach.

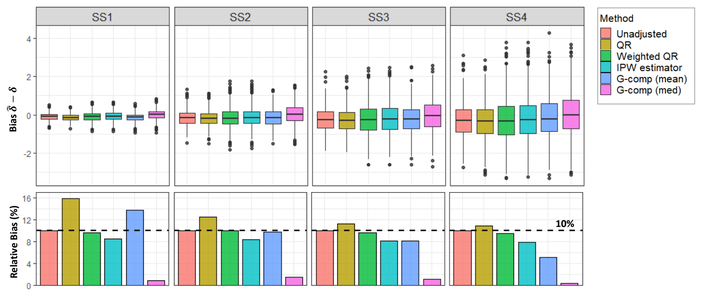

Figure 3. (a) Bias distribution and (b) associated relative bias (%) across 1000 datasets per skewness scenario (SS).

- Estimates were the least biased using the **G-comp (med)** method, with the relative bias substantially reduced.
- Other methods had higher bias or offered little improvement compared to the unadjusted approach.
- Variation in estimates increased for a higher degree of skewness in Y.
- Coverage probability was close to nominal level for all methods and across all skewness scenarios.

Illustrative Example

- We applied the proposed methods to the LSAC study [1], to estimate the CE of maternal mental illness on a child's behaviour in early childhood.
- Estimates were similar for the IPW-based and G-comp methods.
- The QR method estimated a lower CE, although was restricted to the assumption of a linear CE across confounder substrata (not realistic for this application).

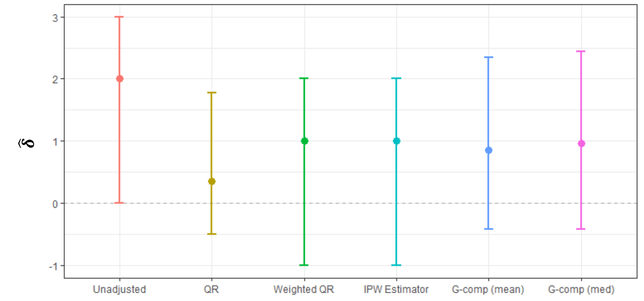

Figure 4. Estimates for δ obtained on the LSAC study under each proposed method, alongside the unadjusted estimate.

Conclusions

- In the presence of skewed outcome data, the common approach to **"ignore or transform"** may not be optimal, and **defining** the causal effect using median potential outcomes may be more appropriate.
- We identified and described a number of confounding-adjustment methods to estimate this causal effect.
- Our simulation study and illustrative example suggest the **G-computation (medians) approach** is the best-performing confounding-adjustment method to estimate this CE using observational data.
- **Future work:** Explore other data generation mechanisms such as those seen in previous studies, and compare consistency of results [3].

References

[1] Sanson A, Nicholson J, Ungerer J, Wilson K, Zubrick S. Introducing the Longitudinal Study of Australian Children. 2002.
[2] Hernán M, A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*. 2004. 58(4):265-271.
[3] Sun S, Moodie E, Neslehova J. Causal inference for quantile treatment effects. *Environmetrics*. 2021. e2668.
[4] Zhang Z, Chen Z, Troendle J, Zhang J. Causal inference on quantiles with an obstetric application. *Biometrics*. 2012. 68:697-706.

Acknowledgements

Funding to present this research was gratefully received from the Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute.