

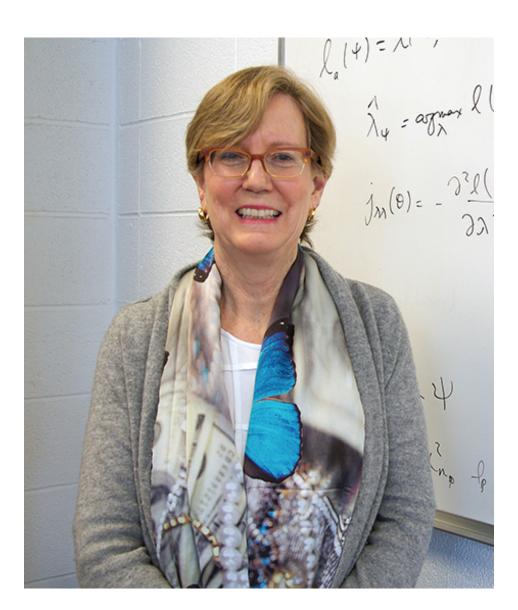


1st INRC

A Semiparametric Approach to Data-Integrated Causal Inference



Archer Gong Zhang
University of Glasgow



Nancy Reid University of Toronto



Qiang Sun U of Toronto & MBZUAI

Causal inference with multi-source data

Goal: estimate the causal effects on a target population.

Collected Data	Experimental (RCT)	Observational (OBS)
Confounding	No	Yes
Representative of the target population	No	Yes
Size	Small	Large
Cost	High	Low
Disadvantage	Lack of external validity	Lack of internal validity

How to take advantage of both types of data with complementary features?

An example of integration:

Based on RCT on *female* and OBS of the real-world usage of the drug on *men*.

U.S. FDA Approves IBRANCE® (palbociclib) for the Treatment of Men with HR+, HER2- Metastatic Breast Cancer

Thursday, April 04, 2019 - 10:57am

Density ratio model (DRM)

- Potential outcome: Y(a) with treatment a.
- Data: $\{(Y_i, X_i, A_i, S_i) : i\}$, where $S_i = 1 (i \in RCT)$.
- Model: for all a = 0, ..., K and s = 0, 1,

user-specified vector-valued functions

$$Y|X,A,S \sim \mathrm{d}G(y|x,a,s) = \exp\{\alpha(x;\theta_{a,s},G_0) + \beta^{\mathsf{T}}(x;\theta_{a,s})q(y)\}\mathrm{d}G_0(y)$$
 "normalizing constant" a common baseline distribution

• Can be seen as a generalization of the GLM: G_0 is unspecified.

Empirical likelihood Inference for Y(a)

Estimate the baseline distribution and model parameters: $\hat{G}_0(y)$ and $\{\hat{\theta}_{a,s}:a,s\}$

- Utilize the entire data to estimate $G_0(y)$.
- Asymptotically efficient $\hat{\theta}_{a,s}$.

Estimate the distribution of $Y(a) \mid X = x$: $\hat{G}(y \mid x, a, s = 1)$ under *Internal Validity*

- Plug-in estimator under the DRM for RCT.
- Consistency & asymptotic normality.

Inference for functionals of Y(a), e.g., mean, CDF, quantiles, under *Transportability*

- Marginalize $\hat{G}(y | x, a, s = 1)$ over the observed x in *OBS*.
- Confidence region & hypothesis test.

imuuchas gracias!

Questions & discussions are welcome! :-)