PROPENSITY SCORE MATCHING AND REWEIGHTING TO ASSESS COMPARATIVE EFFICACY OF SINGLE-ARM TRIALS: A GENERAL FRAMEWORK AND SIMULATION STUDY

Spin P[1], Siddiqui M[2], Hutton B[3], Cameron C[1]

[1]Cornerstone Research Group Inc., Sydney, NS, Canada
[2]Cornerstone Research Group Inc., Burlington, ON, Canada
[3]Ottawa Hospital Research Institute, Ottawa, ON, Canada

INTRODUCTION

While propensity score (PS) matching and reweighting are widely used in many disciplines to balance characteristics across comparison groups, there is a lack of guidance regarding their use with single arm trials, where datasets are typically smaller than those used in observational studies.

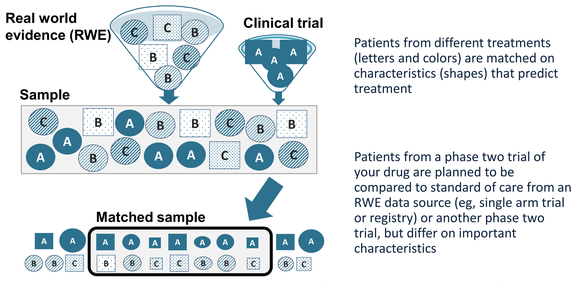
OBJECTIVES

The goal of this analysis is to illustrate a general framework for the use of PS methods for indirect treatment comparisons and to compare the performance of common and emerging PS algorithms using Monte Carlo simulations.

METHODS

A general framework for propensity score analysis with single-arm trials

Conventional Propensity Score Matching



Patients from different treatments (letters and colors) are matched on characteristics (shapes) that predict treatment

Patients from a phase two trial of your drug are planned to be compared to standard of care from an RWE data source (eg, single arm trial or registry) or another phase two trial, but differ on important characteristics

Monte Carlo studies with scenarios defined by treatment arms of varying population sizes and overlap in baseline characteristics.

Scenario	A	B	C	D	E	F
N_A	250	250	300	300	200	200
N_B	250	250	200	200	300	300
$\overline{SMD}_{A-B}$	0.21	0.34	0.21	0.34	0.21	0.34

Propensity score methods evaluated in Monte Carlo simulations

PS Method	Nearest Neighbor Matching (NNM)	Full Optimal Matching (FOM)	Inverse Probability of Treatment Weighting (IPTW)	Covariate Balancing Propensity Score (CBPS)	Entropy Balancing (EB)
	Trial A, Trial B; Match not within caliper; Propensity Score		Propensity Score	Propensity Score $E(Age) = E(Age)$	Age
Description of PS Method	• 1:1 sample, often constructed with a caliper. • Unmatched cases result in loss of power.	• Formulates matched strata using all patients, some in more than one stratum. • May be combined with a caliper and/or weighting.	• Uses all patients. • Weights derived from a logit model for the propensity score. • More weight given to control patients who are similar to treated patients.	• Uses all patients. • Weights derived from a logit model for propensity score that directly equates the moments of baseline covariates between treatment arms.	• Uses all patients. • Weights derived by optimization that directly equates the moments of baseline covariates between treatment arms. • Optimization ensures weights are as close to 1 as possible.

RESULTS

Simulation Results

Scenario	A	B	C	D	E	F
Bias						
NNM Caliper =0.2	20.0	22.8	13.7	18.5	23.5	26.7
IPTW	0.5	6.1	0.8	4.2	0.0	9.5
FOM	8.2	31.3	8.4	29.1	10.2	34.7
EB	-0.1	1.6	0.3	2.5	-1.6	-0.8
CBPS	11.8	35.9	7.0	34.9	16.5	36.8
Coverage						
NNM Caliper =0.2	100	100	100	100	100	100
IPTW	93	62	93	64	89	56
FOM	88	62	89	64	89	58
EB	89	85	89	82	88	82
CBPS	93	60	95	62	88	55
Mean Squared Error (MSE)						
NNM Caliper =0.2	0.023	0.029	0.017	0.026	0.025	0.035
IPTW	0.017	0.086	0.015	0.067	0.021	0.086
FOM	0.021	0.057	0.021	0.053	0.022	0.068
EB	0.021	0.027	0.019	0.030	0.022	0.027
CBPS	0.017	0.059	0.014	0.055	0.021	0.065

Inferior — Superior

Top-ranked PS methods

Scenario	A	B	C	D	E	F
Bias	EB (IPTW)	EB (IPTW)	EB (IPTW)	EB (IPTW)	IPTW (EB)	EB (IPTW)
Coverage	EB (CBPS)	NNM (EB)	CBPS (IPTW)	NNM (EB)	NNM (IPTW + FOM)	NNM (EB)
Mean Squared Error	IPTW (CBPS)	EB (NNM)	CBPS (IPTW)	NNM (EB)	IPTW + CBPS (EB)	EB (NNM)

Heat map of numeric results reported to left for **Bias (smaller is better), coverage (proximity to 95% is better), and Mean Squared Error (smaller is better).** The top-ranked algorithm for each case is reported above, with the second-ranked algorithm listed in parentheses.

The target estimand was the average treatment effect on the treated (ATET). The "true" treatment effect that was used to simulate the data had an ATET of 0.5 on a mean difference scale. Results are based on 1000 simulations.

CONCLUSIONS

Based on these simulations, entropy balancing had superior bias reduction and ranked highly on MSE and coverage. IPTW also yielded favorable bias reduction in all scenarios. NNM had favorable MSE but high bias in most scenarios; its coverage was 100% in all scenarios, suggesting that caliper matching may reduce statistical power when comparing single-arm trials.

ISPOR 2019
May 18-22, 2019 | New Orleans, LA, USA
#ISPORAnnual

Chris Cameron
Cornerstone Research Group Inc.
207-275 Charlotte Street,
Sydney, Nova Scotia, Canada B1P 1C6
Tel: 902.539.4324
E-mail: CCameron@cornerstone-research.com
www.cornerstone-research.com

Cornerstone RESEARCH GROUP