The Korean Journal of Applied Statistics
2023, Vol. 36, No. 5, 381–397

DOI: https://doi.org/10.5351/KJAS.2023.36.5.381
Print ISSN1225-066X / Online ISSN 2383-5818

A simulation study for various propensity score weighting methods in clinical problematic situations

Siseong Jeong[a], Eun Jeong Min[1,a,b]

[a]Department of Biomedicine & Health Sciences, The Catholic University of Korea;
[b]Department of Medical Life Sciences, College of Medicine, The Catholic University of Korea

Abstract

The most representative design used in clinical trials is randomization, which is used to accurately estimate the treatment effect. However, comparison between the treatment group and the control group in an observational study without randomization is biased due to various unadjusted differences, such as characteristics between patients. Propensity score weighting is a widely used method to address these problems and to minimize bias by adjusting those confounding and assess treatment effects. Inverse probability weighting, the most popular method, assigns weights that are proportional to the inverse of the conditional probability of receiving a specific treatment assignment, given observed covariates. However, this method is often suffered by extreme propensity scores, resulting in biased estimates and excessive variance. Several alternative methods including trimming, overlap weights, and matching weights have been proposed to mitigate these issues. In this paper, we conduct a simulation study to compare performance of various propensity score weighting methods under diverse situation, such as limited overlap, misspecified propensity score, and treatment contrary to prediction. From the simulation results overlap weights and matching weights consistently outperform inverse probability weighting and trimming in terms of bias, root mean squared error and coverage probability.

Keywords: propensity score, inverse probability weights, simulation study, limited overlap

1. 서론

최근 부상하고 있는 실사용 증거(real world evidence; RWE)에서 정확한 인과 추론(causal inference)을 하기 위해, 편향(bias)을 일으키는 교란요인(confounding)을 통제하는 다양한 방법들 중 적합한 방법을 선택하는 것이 중요하다. 무작위 대조 시험(randomized controlled trials; RCT)은 인과 추론을 하는 데 있어 치료(treatment) 그룹과 대조(control) 그룹이 유사한 특성을 가지고 있고 나머지 차이가 우연에 의한 것일 가능성을 증가시킴으로써 교란요인을 제어하는 실험계획법이다. RCT결과 그룹 간 차이는 인과 효과(causal effect)에 의해 발생하는 값이며, 추론한 인과관계 추정값은 치료 효과(treatment effect)로 해석될 수 있다. 하지만 치료를 무작위로 배정하는 것은 윤리적 측면에서 문제가 생길 수 있고 결과를 모으는 데 시간이 걸리는 경우가 많아 RCT를 항상 구현할 수는 없다. 이러한 이유로 관찰연구(observational study)에서 인과관계를 추론하는 것이 큰 관심을 받고 있다 (Cochran과 Rubin, 1973). 그러나 관찰연구는 환자 간의 특성(characteristics) 차이로 인해 인과관계 추론에 어려움이 발생한다. 대표적으로 치료와 결과 사이의 관계를 잘못 도출(misleading)하게 만들거나, 실제 효과를 과대평가(overestimate) 또는 과소평가(underestimate)하게 하는 문제점이 있다.

This study was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (No. NRF-2021R1F1A1058613).
[1]Corresponding author: Department of Medical Life Sciences, College of Medicine, The Catholic University of Korea, Banpo-daero 222, Seocho-gu, Seoul 06591, Korea. E-mail: ej.min@catholic.ac.kr

Published 31 October 2023 / journal homepage: http://kjas.or.kr

Table 1: Examples of tilting functions, targeted (sub)populations, causal estimands, and their weights

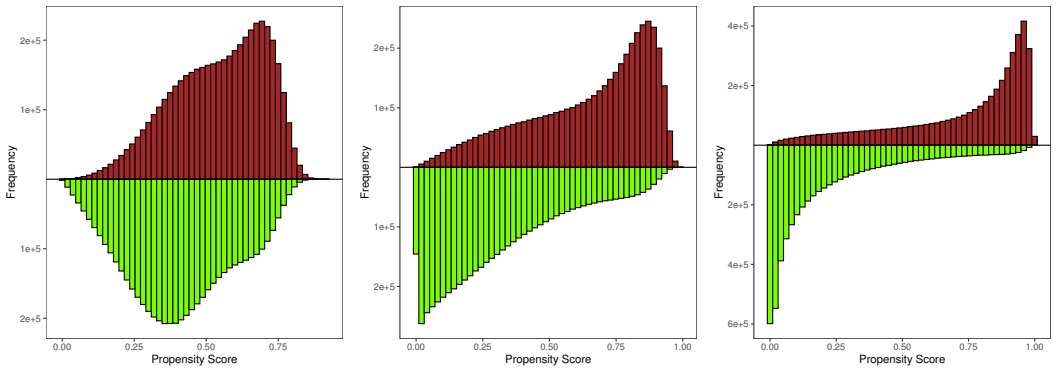
Target population	$h(x)$	Estimand	Weights		Name
			w_1	w_0	
Overall	1	ATE	$e(\boldsymbol{x})^{-1}$	$(1-e(\boldsymbol{x}))^{-1}$	IPW
Trimmed	$\boldsymbol{I}_\alpha(\boldsymbol{x})$	OSATE	$\boldsymbol{I}_\alpha(\boldsymbol{x})e(\boldsymbol{x})^{-1}$	$\boldsymbol{I}_\alpha(\boldsymbol{x})(1-e(\boldsymbol{x}))^{-1}$	trimmed IPW
Treated	$e(\boldsymbol{x})$	ATT	1	$e(\boldsymbol{x})(1-e(\boldsymbol{x}))^{-1}$	IPW for treated
Controls	$1-e(\boldsymbol{x})$	ATC	$e(\boldsymbol{x})(1-e(\boldsymbol{x}))^{-1}$	1	IPW for controls
Overlap	$e(\boldsymbol{x})(1-e(\boldsymbol{x}))$	ATO	$1-e(\boldsymbol{x})$	$e(\boldsymbol{x})$	Overlap weights
Matching	$u(\boldsymbol{x})$	ATM	$u(\boldsymbol{x})e(\boldsymbol{x})^{-1}$	$u(\boldsymbol{x})(1-e(\boldsymbol{x}))^{-1}$	Matching weights

$\boldsymbol{I}_\alpha(\boldsymbol{x}) = \boldsymbol{I}(\alpha \le e(\boldsymbol{x}) \le 1-\alpha)$, where $0 < \alpha < 0.5$ and $\boldsymbol{I}(\cdot)$ is the standard indicator function; $u(\boldsymbol{x}) = \min\{e(\boldsymbol{x}), 1-e(\boldsymbol{x})\}$; OSATE = optimal sample average treatment effect.

관찰연구에서 발생하는 인과 효과 왜곡의 문제를 해결하기 위해 교란요인을 제어하여 인과 효과를 더 정확하게 평가할 수 있도록하는 통계 방법인 성향 점수(propensity score; PS) 방법이 개발되었다 (Rosenbaum과 Rubin, 1983). 성향 점수를 이용하면 그룹 간 관측된 공변량 분포가 균형(balancing)을 이루게 되고 특정 조건에서 무작위 대조 시험 설계의 일부 측면을 모방하여 인과 효과를 적절히 추정할 수 있다. 성향 점수는 피험자의 특성(관측된 공변량)이 주어졌을 때 치료 받을 조건부 확률로, 추정된 성향 점수는 가중치(weighting), 일치(matching) 및 계층화(stratification)를 포함한 다양한 방법으로 교란요인을 조정하여 좀 더 정확한 치료 효과를 추정하도록 한다 (Freedman과 Berk, 2008; Stuart, 2010; Rosenbaum과 Rubin, 1985; Austin, 2008; Lunceford와 Davidian, 2004). 그중에서 본 연구는 각 피험자가 PS의 함수인 w_i에 의해 가중치가 부여되는 성향 점수 가중치를 가지고 치료 그룹과 대조 그룹의 공변량 분포가 균형을 이루도록 한 뒤 인과 효과를 추론한다. 가중치에는 균형 가중치(balancing weights)라고 하는 일반적인 분류(class)가 있으며, 각 분류는 특정 목표 모집단(target-population)을 생성하여 그에 해당하는 가중 평균 처리 효과를 구할 수 있다 (Li 등, 2019).

가장 대표적인 성향 점수 가중치 방법은 역확률 가중치(inverse probability weights; IPW) 방법으로, 관측연구에서 발생하는 그룹 간의 교란요인을 조정하는 데 사용되는 일반적인 접근법이다 (Robins, 1986; Austin과 Stuart, 2015). 역확률 가중치는 관찰된 공변량이 주어졌을 때 특정 치료에 할당될 조건부 확률의 역에 비례하는 가중치를 할당한다. 그러나 실제 분석하고자 하는 데이터에는 일부 환자가 0 또는 1에 가까운 극단적인(extreme) 성향 점수를 보이는 경우가 있는데, 극단적인 성향 점수를 이용한 역확률 가중치는 특정 환자의 가중치를 매우 커지게 만든다. 이런 가중치의 영향을 받은 치료 효과 추정값은 편향되거나 큰 변동성의 문제로 인해 인과 추론의 신뢰성을 저하시킬 수 있다. 최근 대규모 데이터 정보의 보급이 증가함에 따라 극단적인 성향 점수를 처리하기 위한 모범 사례를 명확히 할 필요성이 대두된다 (Li 등, 2019). 이러한 문제를 해결하기 위해 성향 점수가 매우 높거나 낮은 환자를 제외(discard)하는 절사 역확률 가중치(trimmed inverse probability weights; trimmed IPW) 방법이 제안되었다 (Crump 등, 2009). 하지만 절사IPW를 이용한 방법은 종종 불필요하게 너무 많은 환자 정보를 제외한다는 문제점이 존재한다 (Lee 등, 2011). 이를 보완하기 위해 임상적 균형(clinical equipoise)을 이루는 것을 목표로 하는 중복 가중치(overlap weights; OW)와 일치 가중치(matching weights; MW) 방법이 대안으로 제시되었다 (Li 등, 2018; Li와 Greene, 2013). 이 외에도 엔트로피 가중치(entropy weights), 캘리브레이션 가중치(calibration weights) 등 다양한 성향 점수 가중치 방법의 개발 연구가 최근 활발하게 진행되고 있는 추세다 (Zhou 등, 2020; Kim 등, 2017).

본 연구에서는 여러 가지 성향 점수 가중치 방법들을 소개하고, 모의실험을 통해 이들의 성능을 비교하고자 한다. 실제 임상에서 성향 점수를 사용하여 치료 효과를 추정하는 경우 발생할 수 있는 문제상황을 모의실험에 반영하기 위해 우리는 Zhou 등 (2020)과 Stürmer 등 (2021)의 연구를 참고하여 예측에 반하는 치료의 보류가 나타난 경우와 성향 점수 모형을 잘못 지정한 경우를 고려하여 비교한다.



(a) Good overlap (b) Moderate overlap (c) Poor overlap

Figure 1: *Histograms of the propensity scores from simulated data, with good (left), moderate (middle), and poor (right) overlap. Above the horizontal line is the histogram of propensity scores for the treated, and below for the controls.*

본 논문의 구성은 다음과 같다. 2절에서는 먼저 성향 점수의 기본 개념과 평균 치료 효과 추정을 위한 몇 가지 가정을 살펴본 후 3절에서 성향 점수를 이용한 다양한 가중치 방법인 IPW와 절사IPW, OW, MW를 간단히 소개한다. 4절에서는 평균 치료 효과 추정량과 추정 방법에 대해 서술한다. 5절에서는 모의실험을 통해 평균 치료 효과 추정에서 네 가지 가중치 방법들의 성능을 비교한다. 마지막으로 6절에서는 연구결과에 대해 요약하고 향후 연구방향을 제시한다.

2. 성향 점수

성향 점수는 Rosenbaum과 Rubin (1983)에 의해 '관측된 공변량의 벡터가 주어졌을 때 특정 치료에 할당될 조건부 확률'으로 처음 정의되었다. 즉, 성향 점수는 개인의 기저 특성(baseline characteristics)이 정해졌을 때 치료에 노출될(exposed) 조건부 확률로 계산되며, 관찰연구에서 치료 그룹 간 특성이 균형을 이루도록 하여 교란요인을 통제(control)하는 것이 목적이다. 성향 점수가 적용된 피험자 집단은 그룹 간 특성이 균형을 이루는 RCT의 일부를 모방하기에 적절한 치료 효과를 구할 수 있다 (Joffe와 Rosenbaum, 1999).

측정된 치료 전 공변량의 벡터를 $\boldsymbol{X}$, 관찰된 결과는 Y 그리고 치료 할당의 지표 Z를 치료군($Z = 1$)과 대조군($Z = 0$)으로 나타낸다면, 치료 여부가 적용될 관심 모집단 N명의 피험자에게서 관측한 데이터는 $\{(\boldsymbol{X}_i, Z_i, Y_i),\ i = 1, \ldots, N\}$으로 나타낼 수 있다. 이 때, 성향 점수 $e(\boldsymbol{X}_i)$ 는 다음과 같이 정의한다.

$$e(\boldsymbol{X}_i) = P\left(Z_i = z \mid \boldsymbol{X}_i = \boldsymbol{x}\right), \quad \text{for } 0 < e(\boldsymbol{X}) < 1.$$

성향 점수를 이용하여 적절한 치료 효과를 추정하기 위해서는 Rubin (1974)의 잠재적 결과(potential outcome) 체계를 고려해야 한다. i 번째 환자가 치료를 받았을 때의 결과를 $Y_i(1)$, 치료를 받지 않았을 때의 결과를 $Y_i(0)$ 라고 할 때 이 환자에게서 관찰될 수 있는 두 가지 결과 $Y_i(1), Y_i(0)$ 를 잠재적 결과라고 한다. $Y_i(1), Y_i(0)$의 차이나 비율을 이용하여 i 번째 환자에 대한 치료 효과를 나타낼 수 있지만, 실제 연구에서는 두 잠재적 결과를 동시에 관측하는것이 불가능하므로 각 환자의 치료 효과를 구할 수 없다.

반면 관심있는 모집단에 대한 평균 치료 효과는 세 가지 가정을 만족하면 주어진 정보를 이용해 추정할 수 있다 (Kim과 Kim, 2020). 첫 번째는 Rubin (1980)의 안정된 단위 치료값 가정(stable unit treatment value assumption; SUTVA) 이다. 이는 한 환자에게서 관찰 된 결과는 그 환자의 잠재적 결과와 할당된 치료에만

Table 2: The simulation results of homogeneous treatment effect with medium prevalence

Weights	Overlap	True	Unmeasured confounder									
			Complete					Withheld				
			Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
IPW	Good	1	0.47	10.92	10.92	10.03	0.95	3.96	10.21	9.42	8.62	0.91
IPW(0.05)	Good	1	0.46	9.37	9.36	9.08	0.94	3.63	9.15	8.40	8.16	0.91
IPW(0.1)	Good	1	0.44	8.30	8.29	8.24	0.95	3.10	8.24	7.64	7.61	0.93
IPW(0.15)	Good	1	0.40	7.82	7.81	7.76	**0.95**	2.21	7.59	7.26	7.26	0.94
OW	Good	1	0.15	**6.91**	**6.92**	**6.79**	0.94	**0.35**	**6.58**	**6.57**	**6.62**	**0.95**
MW	Good	1	**0.09**	7.22	7.22	7.20	0.95	−0.96	7.56	7.50	7.37	0.94
IPW	Moderate	1	3.05	36.55	36.44	23.01	0.84	9.78	27.16	25.35	15.35	0.70
IPW(0.05)	Moderate	1	0.78	14.58	14.56	14.09	0.94	2.96	15.66	15.38	11.22	0.84
IPW(0.1)	Moderate	1	0.61	10.73	10.72	10.64	0.95	**−0.21**	12.76	12.77	9.59	0.86
IPW(0.15)	Moderate	1	0.20	9.40	9.40	9.45	0.95	−1.20	11.29	11.23	8.97	0.89
OW	Moderate	1	0.11	**7.79**	**7.80**	**7.75**	0.95	−0.23	**7.42**	**7.42**	**7.18**	**0.95**
MW	Moderate	1	**0.11**	8.22	8.22	8.30	**0.96**	−1.18	9.64	9.58	8.13	0.91
IPW	Poor	1	17.13	55.30	52.61	36.14	0.73	10.98	43.41	42.02	21.94	0.66
IPW(0.05)	Poor	1	0.84	16.16	16.14	15.57	0.94	−5.92	24.64	23.93	14.90	0.77
IPW(0.1)	Poor	1	0.31	11.67	11.67	11.65	**0.96**	−7.20	20.58	19.29	12.83	0.82
IPW(0.15)	Poor	1	**0.17**	10.66	10.66	11.02	0.95	−6.80	18.18	16.87	12.09	0.85
OW	Poor	1	0.23	**8.95**	**8.95**	**8.84**	0.95	**−0.05**	**7.89**	**7.90**	**7.73**	**0.94**
MW	Poor	1	0.30	9.42	9.42	9.48	0.95	1.59	10.95	10.84	8.90	0.88
Weights	Overlap	True	Propensity score misspecification									
			Missing X_2					Missing X_6				
			Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
IPW	Good	1	18.23	20.73	9.87	9.46	**0.46**	−3.97	12.04	11.37	10.81	**0.94**
IPW(0.05)	Good	1	18.07	20.19	9.01	9.07	0.47	−4.23	11.22	10.40	10.24	0.93
IPW(0.1)	Good	1	18.06	19.92	8.40	8.58	0.44	−4.35	10.50	9.56	9.65	0.93
IPW(0.15)	Good	1	**18.01**	**19.79**	8.20	8.26	0.42	−4.42	10.33	9.34	9.24	0.93
OW	Good	1	18.66	20.16	**7.63**	**7.59**	0.32	−4.35	9.23	**8.15**	**8.12**	0.91
MW	Good	1	19.57	21.15	8.04	8.01	0.33	**−3.97**	**9.07**	8.16	8.23	0.92
IPW	Moderate	1	29.86	41.20	28.40	18.78	**0.47**	−6.72	36.29	35.68	24.40	0.92
IPW(0.05)	Moderate	1	**28.89**	32.12	14.05	13.77	0.44	−9.07	18.51	16.15	16.04	**0.93**
IPW(0.1)	Moderate	1	29.35	**31.19**	10.57	10.75	0.23	−7.03	13.72	11.78	11.85	0.92
IPW(0.15)	Moderate	1	31.80	33.19	9.51	9.73	0.10	**−5.46**	11.65	10.29	10.46	0.92
OW	Moderate	1	33.84	34.83	**8.27**	**8.39**	0.02	−6.78	11.16	**8.87**	**8.90**	0.88
MW	Moderate	1	36.37	37.40	8.73	8.98	0.01	−5.76	**10.63**	8.94	9.07	0.90
IPW	Poor	1	42.72	58.77	40.39	30.80	**0.53**	**−2.21**	57.87	57.86	41.76	0.85
IPW(0.05)	Poor	1	**37.16**	**40.17**	15.29	15.20	0.33	−9.57	19.47	16.97	16.72	0.92
IPW(0.1)	Poor	1	42.49	43.98	11.37	11.48	0.04	−6.88	14.26	12.49	12.52	0.92
IPW(0.15)	Poor	1	46.85	48.12	10.98	10.93	0.01	−5.22	12.76	11.64	11.75	**0.93**
OW	Poor	1	46.54	47.44	**9.20**	**9.24**	0.00	−7.72	12.56	**9.91**	**9.83**	0.86
MW	Poor	1	49.73	50.66	9.65	9.87	0.00	−6.44	**11.98**	10.11	10.11	0.89

Bias = relative bias×100; RMSE = root mean-squared error×100; SD = empirical standard deviation×100; SE = average estimated standard error×100; CP = 95% coverage probability; IPW(α) = trimmed IPW with $\boldsymbol{I}_\alpha(\boldsymbol{x}) = \boldsymbol{I}(\alpha \le e(\boldsymbol{x}) \le 1-\alpha)$, for $\alpha = 0.05, 0.1, 0.15$.

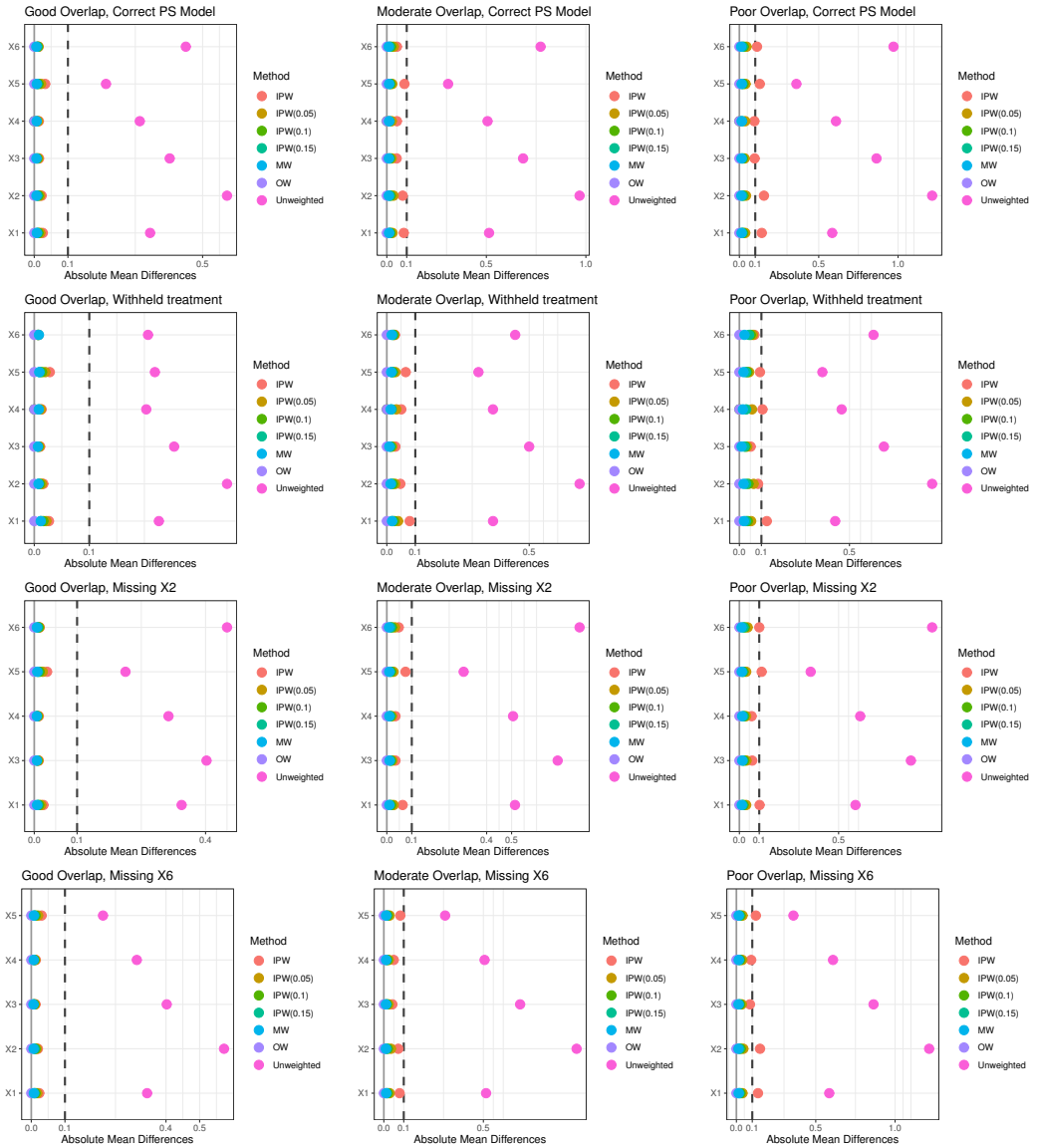


Figure 2: *Covariate balance with absolute standardized mean difference.*

영향을 받는다는 것으로, $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$을 만족한다. 두 번째는 치료할당의 조건부 독립성으로, 공변량 $\boldsymbol{X}$가 주어졌을 때 치료 할당은 잠재적 결과와 독립이라는 것이다. 교란요인을 제어하기에 충분한 공변량 $\boldsymbol{X}$는 $E[Y(z)|Z = z, \boldsymbol{X}] = E[Y(z)|\boldsymbol{X}]$ $(z = 0, 1)$ 을 만족한다. 마지막 세 번째는 성향 점수를 이용한 방법을 고려할 때 가장 중요한 가정인 양수성(positivity) 가정이다. 이것은 모든 환자에대해 치료를 받을 확률은 0과 1사이에 존재해야 한다는 것으로 $P(\{\boldsymbol{x} : v < e(\boldsymbol{x}) < 1 - v\}) = 1$ $(v > 0)$이 만족되어야 함을 의미한다 (Rosenbaum과 Rubin, 1983).

관찰연구에서 참 성향 점수는 알 수 없으므로 추정하여 구하는데, 이진(binary) 치료의 경우 성향 점수를

Table 3: The simulation results of heterogeneous treatment effect with medium prevalence

Weights	Overlap	True	Unmeasured confounder									
			None					Withheld				
			Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
IPW	Good	1.53	0.35	11.72	11.72	10.60	0.94	3.73	11.69	10.22	9.01	0.87
IPW(0.05)	Good	1.53	0.38	9.81	9.80	9.48	0.94	3.41	10.39	9.00	8.48	0.87
IPW(0.1)	Good	1.54	0.37	8.54	8.52	8.45	0.94	2.85	9.13	8.01	7.81	0.91
IPW(0.15)	Good	1.55	0.35	7.94	7.93	7.86	0.95	2.03	8.17	7.54	7.37	0.92
OW	Good	1.55	0.11	**6.94**	**6.94**	**6.83**	0.94	0.66	**6.67**	**6.59**	**6.64**	**0.95**
MW	Good	1.57	**0.02**	7.21	7.21	7.19	**0.95**	**−0.56**	7.50	7.46	7.37	0.94
IPW	Moderate	1.33	2.66	40.08	39.94	24.81	0.81	11.43	32.16	28.37	16.55	0.61
IPW(0.05)	Moderate	1.37	0.89	15.22	15.18	14.59	0.94	4.33	17.98	16.99	11.80	0.78
IPW(0.1)	Moderate	1.42	0.65	11.04	11.01	10.85	0.94	**−0.29**	13.76	13.77	9.88	0.84
IPW(0.15)	Moderate	1.48	0.17	9.61	9.61	9.56	0.95	−3.40	12.88	11.86	9.12	0.85
OW	Moderate	1.44	0.14	**7.87**	**7.88**	**7.88**	0.95	−0.64	**7.64**	**7.58**	**7.28**	**0.95**
MW	Moderate	1.47	**0.05**	8.23	8.23	8.36	**0.96**	−2.24	10.41	9.87	8.24	0.89
IPW	Poor	1.17	16.00	59.41	56.39	38.12	0.71	16.65	49.91	45.95	23.17	0.59
IPW(0.05)	Poor	1.30	1.09	16.60	16.54	16.04	0.94	−5.38	26.45	25.53	15.42	0.74
IPW(0.1)	Poor	1.40	0.38	12.05	12.05	11.87	0.95	−10.80	25.29	20.31	13.11	0.71
IPW(0.15)	Poor	1.47	**0.04**	10.87	10.88	11.14	**0.96**	−12.79	26.20	18.24	12.26	0.64
OW	Poor	1.38	0.26	**9.08**	**9.08**	**9.04**	0.95	−3.29	**9.43**	**8.26**	**7.90**	**0.89**
MW	Poor	1.43	0.18	9.50	9.50	9.61	0.95	**−3.26**	12.97	12.11	9.16	0.84
Weights	Overlap	True	Propensity score misspecification									
			Missing X_2					Missing X_6				
			Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
IPW	Good	1.53	12.79	22.14	10.42	9.84	0.44	**−2.35**	12.41	11.88	11.14	**0.94**
IPW(0.05)	Good	1.53	12.58	21.39	9.34	9.37	**0.45**	−2.56	11.34	10.65	10.45	0.93
IPW(0.1)	Good	1.54	12.28	20.74	8.59	8.78	0.41	−2.70	10.49	9.64	9.71	0.93
IPW(0.15)	Good	1.55	**11.88**	**20.19**	8.34	8.40	0.42	−2.77	10.28	9.35	9.22	0.93
OW	Good	1.55	12.20	20.46	**7.72**	**7.68**	0.33	−2.71	**9.11**	**8.08**	**8.05**	0.91
MW	Good	1.57	12.35	20.99	8.10	8.08	0.35	−2.50	8.98	8.09	8.16	0.92
IPW	Moderate	1.33	25.99	46.10	30.63	19.83	**0.41**	−3.95	38.67	38.33	25.68	0.91
IPW(0.05)	Moderate	1.37	22.89	34.49	14.51	14.13	0.40	−6.05	18.50	16.56	16.33	**0.94**
IPW(0.1)	Moderate	1.42	21.13	**31.90**	10.90	10.97	0.23	−4.39	13.49	11.98	11.94	0.93
IPW(0.15)	Moderate	1.48	**20.59**	31.98	9.81	9.88	0.13	**−3.53**	11.67	10.44	10.49	0.93
OW	Moderate	1.44	23.44	34.80	**8.42**	**8.57**	0.02	−4.40	10.87	**8.83**	**8.90**	0.89
MW	Moderate	1.47	23.67	36.01	8.82	9.11	0.02	−3.73	**10.45**	8.88	9.05	0.91
IPW	Poor	1.17	42.92	65.80	42.35	31.82	**0.48**	**−2.23**	64.87	64.85	44.09	0.81
IPW(0.05)	Poor	1.30	30.20	42.21	15.65	15.55	0.31	−6.05	19.48	17.84	17.06	**0.93**
IPW(0.1)	Poor	1.40	28.88	**42.03**	11.72	11.67	0.07	−4.85	14.33	12.63	12.71	0.93
IPW(0.15)	Poor	1.47	**27.56**	42.04	11.11	11.00	0.04	−4.00	13.36	12.00	11.84	0.92
OW	Poor	1.38	32.29	45.63	**9.34**	**9.43**	0.00	−5.39	12.31	**9.80**	**9.95**	0.89
MW	Poor	1.43	31.93	46.61	9.73	9.97	0.00	−4.61	**12.04**	10.08	10.18	0.90

Bias = relative bias×100; RMSE = root mean-squared error×100; SD = empirical standard deviation×100; SE = average estimated standard error×100; CP = 95% coverage probability; IPW(α) = trimmed IPW with $\boldsymbol{I}_\alpha(\boldsymbol{x}) = \boldsymbol{I}(\alpha \le e(\boldsymbol{x}) \le 1-\alpha)$, for $\alpha = 0.05, 0.1, 0.15$.

Table 4: The simulation results of homogeneous treatment effect with low prevalence

Weights	Overlap	True	Unmeasured confounder									
			Complete					Withheld				
			Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
IPW	Good	1	0.55	18.06	18.06	15.76	0.93	11.97	19.38	15.25	13.23	0.77
IPW(0.05)	Good	1	0.13	12.59	12.60	12.01	0.94	10.03	15.68	12.06	11.09	0.83
IPW(0.1)	Good	1	−0.09	10.21	10.21	9.91	0.95	7.48	12.81	10.40	9.70	0.87
IPW(0.15)	Good	1	−0.17	9.32	9.32	9.10	0.94	4.77	10.61	9.48	9.25	0.93
OW	Good	1	−0.07	7.83	7.83	7.50	0.95	0.12	7.85	7.86	7.58	0.95
MW	Good	1	−0.04	8.43	8.43	8.39	0.96	−4.83	9.99	8.75	8.34	0.90
IPW	Moderate	1	17.21	61.21	58.77	37.08	0.71	51.54	70.03	47.43	29.61	0.42
IPW(0.05)	Moderate	1	−0.30	16.34	16.34	16.04	0.94	20.39	27.93	19.09	15.84	0.68
IPW(0.1)	Moderate	1	0.30	12.66	12.66	12.85	0.94	6.72	15.47	13.94	13.07	0.90
IPW(0.15)	Moderate	1	−0.11	12.52	12.53	12.37	0.95	−2.02	16.39	16.27	13.34	0.92
OW	Moderate	1	−0.28	9.42	9.42	9.43	0.96	0.12	9.78	9.78	9.63	0.95
MW	Moderate	1	−0.32	10.51	10.51	10.60	0.95	−7.70	13.36	10.93	10.72	0.87
IPW	Poor	1	58.23	111.74	95.41	50.40	0.49	106.23	131.72	77.93	41.43	0.27
IPW(0.05)	Poor	1	1.18	18.75	18.72	17.82	0.93	15.22	24.18	18.80	17.12	0.82
IPW(0.1)	Poor	1	0.43	15.59	15.59	15.32	0.95	1.11	18.26	18.23	15.24	0.91
IPW(0.15)	Poor	1	−0.08	15.04	15.05	14.90	0.95	−1.06	30.54	30.54	18.11	0.92
OW	Poor	1	−0.14	11.02	11.03	11.19	0.95	0.55	11.70	11.69	11.56	0.95
MW	Poor	1	−0.25	12.21	12.21	12.48	0.96	−7.09	14.90	13.11	12.82	0.91
True	Overlap	True	Propensity score misspecification									
			Missing X_2					Missing X_6				
			Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
IPW	Good	1	19.57	25.54	16.42	14.20	0.61	−0.93	17.58	17.56	15.44	0.93
IPW(0.05)	Good	1	19.40	23.05	12.44	12.09	0.60	−1.98	13.31	13.17	12.42	0.93
IPW(0.1)	Good	1	18.70	21.59	10.79	10.49	0.54	−2.46	10.78	10.50	10.24	0.94
IPW(0.15)	Good	1	18.02	20.67	10.14	9.80	0.54	−2.19	9.73	9.49	9.41	0.94
OW	Good	1	16.18	18.36	8.67	8.30	0.48	−4.21	10.00	9.08	8.84	0.92
MW	Good	1	13.52	16.28	9.08	8.95	0.67	−5.23	11.42	10.15	10.26	0.92
IPW	Moderate	1	43.39	63.54	46.45	32.85	0.53	18.22	57.78	54.86	37.78	0.73
IPW(0.05)	Moderate	1	35.04	38.63	16.27	16.20	0.44	−0.34	15.72	15.72	15.60	0.94
IPW(0.1)	Moderate	1	29.02	32.55	14.74	13.70	0.44	−0.71	12.40	12.39	12.70	0.95
IPW(0.15)	Moderate	1	12.80	19.88	15.21	13.16	0.79	−1.41	12.62	12.55	12.38	0.95
OW	Moderate	1	23.06	25.31	10.43	10.20	0.37	−5.76	11.90	10.43	10.56	0.92
MW	Moderate	1	17.64	20.92	11.24	11.10	0.64	−7.36	14.04	11.96	12.21	0.90
IPW	Poor	1	72.68	113.71	87.49	48.17	0.43	57.91	114.94	99.33	53.56	0.51
IPW(0.05)	Poor	1	44.08	48.38	19.93	18.36	0.35	1.72	18.14	18.06	17.51	0.94
IPW(0.1)	Poor	1	16.90	24.95	18.36	15.66	0.77	−0.43	14.72	14.72	15.02	0.95
IPW(0.15)	Poor	1	9.11	18.46	16.06	15.47	0.91	−1.58	14.61	14.54	14.76	0.95
OW	Poor	1	25.76	28.49	12.17	12.00	0.43	−5.63	13.03	11.76	12.09	0.92
MW	Poor	1	19.01	23.00	12.96	12.90	0.68	−7.46	15.36	13.44	13.81	0.92

Bias = relative bias×100; RMSE = root mean-squared error×100; SD = empirical standard deviation×100; SE = average estimated standard error×100; CP = 95% coverage probability; IPW(α) = trimmed IPW with $\boldsymbol{I}_\alpha(\boldsymbol{x}) = \boldsymbol{I}(\alpha \le e(\boldsymbol{x}) \le 1-\alpha)$, for $\alpha = 0.05, 0.1, 0.15$.

Table 5: The simulation results of geterogeneous treatment effect with low prevalence

Weights	Overlap	True	Unmeasured confounder									
			Complete					Withheld				
			Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
IPW	Good	1.38	0.47	19.46	19.46	16.90	0.92	9.90	21.36	16.39	14.03	0.74
IPW(0.05)	Good	1.40	0.32	13.33	13.33	12.66	0.94	8.01	16.97	12.74	11.61	0.81
IPW(0.1)	Good	1.44	0.20	10.70	10.70	10.26	0.94	5.50	13.35	10.76	10.00	0.86
IPW(0.15)	Good	1.49	−0.08	9.64	9.65	9.29	0.94	3.02	10.74	9.76	9.44	0.92
OW	Good	1.47	−0.03	7.96	7.96	7.57	0.95	−0.16	7.92	7.92	7.65	0.95
MW	Good	1.50	−0.10	8.50	8.50	8.40	0.96	4.18	10.79	8.80	8.28	0.86
IPW	Moderate	1.13	16.85	66.52	63.78	40.17	0.70	51.26	77.41	51.52	31.91	0.42
IPW(0.05)	Moderate	1.30	0.08	17.67	17.67	17.11	0.93	16.17	29.10	20.08	16.81	0.70
IPW(0.1)	Moderate	1.42	0.26	13.34	13.35	13.36	0.94	5.35	16.02	14.12	13.54	0.90
IPW(0.15)	Moderate	1.49	−0.21	12.89	12.89	12.60	0.95	−0.09	15.96	15.97	13.50	0.92
OW	Moderate	1.38	−0.17	9.71	9.72	9.65	0.96	0.77	−10.20	10.15	9.88	0.94
MW	Moderate	1.43	−0.32	10.74	10.73	10.72	0.95	−7.39	15.75	11.67	10.76	0.81
IPW	Poor	1.01	61.62	120.18	102.77	54.71	0.50	115.26	143.90	84.22	44.85	0.27
IPW(0.05)	Poor	1.26	6.19	21.51	20.06	18.99	0.89	18.55	30.59	19.84	18.19	0.72
IPW(0.1)	Poor	1.34	6.81	18.77	16.41	15.86	0.89	8.88	21.96	18.47	15.66	0.83
IPW(0.15)	Poor	1.39	7.16	18.40	15.49	15.15	0.88	7.81	31.46	29.54	18.19	0.84
OW	Poor	1.31	5.09	13.22	11.44	11.48	0.92	4.18	13.42	12.27	11.89	0.91
MW	Poor	1.35	5.45	14.54	12.57	12.68	0.92	−2.35	14.55	14.21	12.96	0.91
Weights	Overlap	True	Propensity score misspecification									
			Missing X_2					Missing X_6				
			Bias	RMSE	SD	SE	CP	Bias	RMSE	SD	SE	CP
IPW	Good	1.38	16.33	28.64	17.59	15.18	0.57	−0.27	18.79	18.79	16.42	0.93
IPW(0.05)	Good	1.40	15.56	25.49	13.21	12.81	0.58	−1.15	13.96	13.87	13.00	0.93
IPW(0.1)	Good	1.44	13.63	22.67	11.37	11.02	0.54	−1.57	11.13	10.91	10.50	0.94
IPW(0.15)	Good	1.49	11.65	20.33	10.61	10.21	0.59	−1.42	10.04	9.82	9.52	0.94
OW	Good	1.47	11.43	18.99	8.95	8.56	0.48	−2.86	10.04	9.12	8.79	0.92
MW	Good	1.50	9.00	16.34	9.25	9.07	0.67	−3.60	11.50	10.16	10.19	0.92
IPW	Moderate	1.13	45.93	72.14	50.24	35.61	0.51	18.67	62.70	59.09	40.78	0.72
IPW(0.05)	Moderate	1.30	28.34	40.98	17.78	17.46	0.45	0.12	17.10	17.11	16.62	0.94
IPW(0.1)	Moderate	1.42	20.30	32.58	15.23	14.50	0.49	−0.24	12.97	12.98	13.11	0.95
IPW(0.15)	Moderate	1.49	8.94	19.97	14.86	13.55	0.80	−0.78	12.89	12.84	12.54	0.94
OW	Moderate	1.38	17.24	26.21	10.91	10.60	0.38	−4.15	12.02	10.57	10.64	0.92
MW	Moderate	1.43	12.29	21.04	11.54	11.29	0.64	−5.31	14.26	12.07	12.25	0.90
IPW	Poor	1.01	59.03	62.30	17.59	15.18	0.09	62.14	123.67	106.52	58.08	0.51
IPW(0.05)	Poor	1.26	28.93	38.64	13.21	12.81	0.22	6.81	21.16	19.37	18.61	0.89
IPW(0.1)	Poor	1.34	22.14	31.74	11.37	11.02	0.26	6.49	17.65	15.37	15.43	0.90
IPW(0.15)	Poor	1.39	19.63	29.26	10.61	10.21	0.24	6.49	17.37	14.85	14.93	0.90
OW	Poor	1.31	25.18	34.06	8.95	8.56	0.03	0.87	12.06	12.02	12.27	0.95
MW	Poor	1.35	21.25	30.06	9.25	9.07	0.13	0.03	13.66	13.67	13.95	0.96

Bias = relative bias×100; RMSE = root mean-squared error×100; SD = empirical standard deviation×100; SE = average estimated standard error×100; CP = 95% coverage probability; IPW(α) = trimmed IPW with $\boldsymbol{I}_\alpha(\boldsymbol{x}) = \boldsymbol{I}(\alpha \le e(\boldsymbol{x}) \le 1-\alpha)$, for $\alpha = 0.05, 0.1, 0.15$.

추정하는 방법에는 판별분석(discriminant analysis)의 사용, 프로빗 회귀 모형(probit regression)의 사용 등이 있으며 가장 널리 알려져 있는 방법은 로지스틱 회귀 모형(logistic regression model)을 통한 방법이다. 로지스틱 회귀 모형의 경우 성향 점수를 $e(\boldsymbol{X}_i;\boldsymbol{\beta}) = 1/(1+\exp(-\boldsymbol{X}_i^T\boldsymbol{\beta}))$ 의 값으로 추정할 수 있다. 추정된 성향 점수는 가중치, 일치, 계층화 등 여러 가지 방법을 통해 공변량을 조정하여 적절한 치료 효과를 얻게 한다.

3. 다양한 성향 점수 가중치 방법

3.1. 역확률 가중치

역확률 가중치(IPW)는 성향 점수 가중치 방법 중 가장 널리 쓰이며, 치료 그룹으로 할당될 확률의 역수에 비례하는 가중치를 각 환자에게 부여하여 기저 특성의 균형을 맞추는 방법으로 다음과 같이 나타낸다 (Austin과 Stuart, 2015).

$$\begin{cases} w_1(\boldsymbol{x}) \propto \dfrac{1}{e(\boldsymbol{x})}, & z = 1, \\ w_0(\boldsymbol{x}) \propto \dfrac{1}{1-e(\boldsymbol{x})}, & z = 0. \end{cases}$$

$\hat{e}_i$는 추정된 성향 점수를 나타내며, 치료를 받은 환자의 가중치는 $\hat{w}_i = 1/\hat{e}_i$, 대조군에 존재하는 환자의 가중치는 $\hat{w}_i = 1/(1-\hat{e}_i)$이다. 하지만 이 방법은 성향 점수가 0 또는 1에 가까운 극단적인 값을 갖는 경우 이 가중치의 값이 매우 크거나 작게 된다. 치료 그룹간의 균형이 잘 이루어지지 않고 평균 치료 효과를 추정하는 것에 있어서 편향되고 불안정(unstable)하며, 매우 큰 분산을 갖게 하는 문제점이 발생한다.

3.2. 절사 역확률 가중치

Crump 등 (2009)은 역확률 가중치의 극단적인 성향 점수로 인해 발생하는 문제점을 해결하기 위해 임계값(threshold) $\boldsymbol{\alpha}$를 지정하여 $[\boldsymbol{\alpha}, 1-\boldsymbol{\alpha}]$를 벗어나는 범위의 성향 점수를 잘라내는 대칭 절사(symmetric trimming) 방법을 제안했다. 임계값 α는 임상 전문가의 지식과 견해에 따라 결정 되며, 이 방법을 절사 역확률 가중치(trimmed IPW)라고 한다.

예를 들어 $\boldsymbol{\alpha} = 0.1$ 인 경우 $[0.1, 0.9]$를 벗어나는 피험자를 제외 한 후 역확률 가중치 방법으로 평균 치료 효과를 추정하게 된다. 이 방법을 사용하면 $[\boldsymbol{\alpha}, 1-\boldsymbol{\alpha}]$값 범위를 벗어나는 점수를 손쉽게 잘라내어 극단적인 성향 점수로 인해 생기는 편향 등의 문제를 빠르게 해결할 수 있지만, 임상적 지침이 없는 $\boldsymbol{\alpha}$값에 따라 결과값이 매우 민감(sensitive)하고 절사를 통해 성향 점수를 제거하면 많은 양의 환자 정보를 잃게 된다 (Lee 등, 2011). 또한 성향 점수에 기반한 임계값 기준은 임상적으로 해석하기가 어렵다는 한계점이 있다 (Traskin과 Small, 2011).

3.3. 중복 가중치

중복 가중치(OW)는 IPW 및 절사IPW에서 발생하는 일부 문제를 해결하기 위해 Li 등 (2018)이 제안한 균형 가중치 체계이다. OW는 각각의 피험자에게 상대 그룹에 할당 될 확률로 가중치를 주는 방식이며, 다음과 같이 정의된다.

$$\begin{cases} w_1(\boldsymbol{x}) \propto 1-e(\boldsymbol{x}), & z = 1, \\ w_0(\boldsymbol{x}) \propto e(\boldsymbol{x}), & z = 0. \end{cases}$$

성향 점수가 0.5 근처에 있는 피험자는 치료 효과의 추정에 기여도(contribution)를 높이기 위해 큰 가중치를 부여하고, 성향 점수가 0 또는 1에 가까운 값을 갖는 PS 분포 꼬리(tail)에 위치한 피험자에게는 작은 가중치

를 부여하는 방법이다. OW의 목표 모집단은 중복되는 특성을 갖고있는 피험자를 강조하는데, 이렇게 OW를 적용하여 재분포한 모집단을 중복 모집단이라고 표현한다. 또한 이 모집단의 효과 추정값은 중복 모집단의 평균 치료 효과(average treatment effect for the overlap population; ATO)이다. OW는 성향 점수가 극단적일 때 이 전의 방법들에 비해 치료 효과를 더 정확하게 추정하므로, IPW가 가졌던 불확실성의 문제를 해결한다. 실제로 OW는 0과 1 사이의 값을 가지며, 모든 균형 가중치 중에서 OW가 치료 효과 추정값의 표본 분산을 최소화한다. 특히 로지스틱 PS 모형을 이용하여 추정된 ATO의 분산은 일관적인(consistent) 추정량을 가지며, 이에 기반한 OW는 모든 공변량에 대한 치료 그룹 사이의 정확한(exact) 균형을 이끈다 (Li 등, 2019; Austin과 Stuart, 2015).

3.4. 일치 가중치

Li와 Greene (2013)는 1 : 1매칭(pair matching)과 IPW의 단점을 보완한 일치 가중치(MW)를 제안했다. 일치(matching)된 피험자와 일치되지 않은 피험자를 분류(classification)한 뒤, 일치되지 않은 값은 제외하는 Rubin (1973)이 제안한 1 : 1매칭 방법과는 다르게 이 방법은 극단적인 성향 점수를 가진 피험자가 일치 되지 않더라도 분석에서 제외시키지 않고 작은 가중치를 부여한다. MW는 다음과 같이 나타낸다.

$$\begin{cases} w_1(\boldsymbol{x}) \propto \min\left(1-e(\boldsymbol{x}), e(\boldsymbol{x})\right)\ e(\boldsymbol{x})^{-1}, & z=1, \\ w_0(\boldsymbol{x}) \propto \min\left(1-e(\boldsymbol{x}), e(\boldsymbol{x})\right)\ (1-e(\boldsymbol{x}))^{-1}, & z=0. \end{cases}$$

MW는 분자에 1 대신 $\min(1-e(\boldsymbol{x}), e(\boldsymbol{x}))$이 위치함으로써 IPW와는 상이한 해석과 수치적 특성을 갖는다. 특정 성향 점수 $\tilde{e}$의 주변에 존재하는 $\tilde{m}$명의 피험자에 대해 대략적으로 $\tilde{m}\tilde{e}$명의 피험자가 치료군에 속하고, $\tilde{m}(1-\tilde{e})$명은 대조군에 배정된 것으로 예상한다. 예를 들어 $\tilde{e} > 0.5$인 경우 성향 점수 $\tilde{e}$의 주변은 대조군보다 치료군에 속하는 피험자가 더 많기에 대조군에 속하는 모든 피험자를 일치된 자료로 선택할 수 있게 된다. 이 때 일치된 자료로 선택될 확률이 대조군 피험자는 1이고 치료군은 $(1-\tilde{e})/\tilde{e}$이다. 쉽게 말해 MW는 각 환자에게 일치된 자료로 선택될 확률로 가중치를 부여하는 방법이라고 할 수 있으며, 잠재적으로 전체가 일치된다고 판단되는 때의 가중치는 모든 피험자가 추정에 기여할 수 있도록 균등한(equal) 분포를 이룬다. MW를 이용하여 구한 효과 추정값은 일치 모집단의 평균 치료 효과(average treatment effect for the matching population; ATM)로 표현한다. 이 가중치를 통해 피험자간 균형이 개선되고 평균 치료 효과의 편향이 감소하며 계산이 안정화(stabilize)되는 것을 확인 할 수 있다 (Li와 Greene, 2013).

4. 가중 평균 치료 효과

4.1. 치료 효과 목표 추정값

앞서 설명한 다양한 가중치들은 치료 그룹과 대조 그룹 간의 공변량의 분포가 균형을 이룬 목표 모집단을 생성하기 위해 사용된다. 함수 $f(\boldsymbol{x})$는 공변량 $\boldsymbol{X}$의 주변밀도함수이며 특정 목표 모집단을 생성하기 위한 지정함수(titling function) $h(\boldsymbol{x})$를 이용하여 생성한 목표 모집단을 함수 $f(\boldsymbol{x})h(\boldsymbol{x})$로 나타낸다. 이 목표 모집단은 $f(\boldsymbol{x})$에서 추출된 환자의 특성이 임상적, 통계적으로 균형을 이룬 분포로 재분포한 결과모집단이다. 치료 그룹 z에 속한 $\boldsymbol{X}$의 확률밀도함수는

$$\begin{aligned} &f_z(\boldsymbol{x}) = \Pr\left(\boldsymbol{X}=\boldsymbol{x} \mid Z=z\right) \text{ 일 때,} \\ &\quad f_1(\boldsymbol{x}) \propto f(\boldsymbol{x})e(\boldsymbol{x}), && z=1, \\ &\quad f_0(\boldsymbol{x}) \propto f(\boldsymbol{x})\left(1-e(\boldsymbol{x})\right), && z=0 \end{aligned}$$

을 만족한다. 또한 지정 함수 $h(\boldsymbol{x})$가 주어졌을 때, 그에 상응하는 그룹별 균형 가중치 $w_z(\boldsymbol{x})$는

$$\begin{cases} w_1(\boldsymbol{x}) \propto \dfrac{f(\boldsymbol{x})h(\boldsymbol{x})}{f(\boldsymbol{x})e(\boldsymbol{x})} = \dfrac{h(\boldsymbol{x})}{e(\boldsymbol{x})}, & z = 1, \\ w_0(\boldsymbol{x}) \propto \dfrac{f(\boldsymbol{x})h(\boldsymbol{x})}{f(\boldsymbol{x})\,(1-e(\boldsymbol{x}))} = \dfrac{h(\boldsymbol{x})}{(1-e(\boldsymbol{x}))}, & z = 0 \end{cases} \tag{4.1}$$

로 정의된다 (Li 등, 2018). 가중 공변량 분포의 균형을 맞추기 위해 균형 가중치 $w_z(\boldsymbol{x})$는 $f_1(\boldsymbol{x})w_1(\boldsymbol{x}) = f_0(\boldsymbol{x})w_0(\boldsymbol{x}) = f(\boldsymbol{x})h(\boldsymbol{x})$ 목표 모집단을 만족한다. 이상적인 목표 모집단은 서로 다른 임상 상황과 초기 피험자 모집 방식 및 통계적 고려사항으로 인해 의학 연구에 따라 달라질 수 있다 (Mao 등, 2019).

함수 $h(\boldsymbol{x})$ 를 통해 가중 평균 치료 효과(WATE)는 다음과 같이 나타낼 수 있다 (Hirano 등, 2003).

$$\Delta_h = \frac{E\left[h(\boldsymbol{X})\left(Y(1)-Y(0)\right)\right]}{E\left(h(\boldsymbol{X})\right)}. \tag{4.2}$$

식 (4.2)의 Δ_h는 추정된 성향 점수가 아닌 실제(true) 성향 점수로 정의된다. Table 1에서 보이는 바와 같이 $h(\boldsymbol{x})$값이 $1, e(\boldsymbol{x}), 1-e(\boldsymbol{x})$ 인 경우 이에 해당하는 각각의 치료 효과(estimands) Δ_h는 평균 치료 효과(average treatment effect; ATE), 치료군에 대한 평균 치료 효과(average treatment treat; ATT), 대조군에 대한 평균 치료 효과(average treatment control; ATC)에 해당한다. 처리 효과가 동질적일 경우 모든 $h(\boldsymbol{x})$에 대해 치료 효과 추정치가 동일하지만, 처리 효과가 이질적인 경우에는 지시 함수 $h(\boldsymbol{x})$에 따라 다른 추정치로 이어질 수 있다. 식 (4.2)의 가중 평균 치료 효과 Δ_h는 잠재적 결과 $Y(z)$와 실제 성향 점수 $e(\boldsymbol{X})$의 존재를 가정하기 때문에 정의될 수 있으며, 추정값의 식별가능성(identifiability)은 SUTVA와 치료 할당의 조건부 독립성에 의해 보장된다.

4.2. 치료 효과의 추정 방법

가중 평균 치료 효과 Δ_h의 추정치 $\hat{\Delta}_h$은 다음과 같이 추정(estimation)할 수 있다.

$$\hat{\Delta}_h = \frac{\sum_{i=1}^{N} Z_i Y_i \hat{w}_i(\boldsymbol{x})}{\sum_{i=1}^{N} Z_i \hat{w}_i(\boldsymbol{x})} - \frac{\sum_{i=1}^{N} (1-Z_i)\, Y_i \hat{w}_i(\boldsymbol{x})}{\sum_{i=1}^{N} (1-Z_i)\, \hat{w}_i(\boldsymbol{x})}, \tag{4.3}$$

여기서 $\hat{w}(\boldsymbol{x}) = Z\hat{w}_1(\boldsymbol{x}) + (1-Z)\hat{w}_0(\boldsymbol{x})$는 추정 가중치이며, 대부분의 관찰 연구에서 PS는 알려져있지 않기에 다양한 PS추정 방법을 이용하여 구한다. 식 (4.2)에서 $h(\boldsymbol{x}) = 1$인 경우의 추정치 $\hat{\Delta}_h$는 수정된 Horvitz-Thompson IPW 추정량(Hajek's 추정량)에 해당한다 (Godambe, 1970). 이 Hejek's estimator를 이용하여 ATO (average treatment effect of overlap population), ATM (average treatment effect of matching population)은 ATE와 동일한 추정치를 추정한다고 볼 수 있다 (Mao 등, 2019; Austin, 2022).

5. 모의실험

5.1. 모의실험 목적

본 절에서는 다양한 가중치 방법이 치료군과 대조군의 비율인 치료분율(average treatment prevalence), 개인별 치료 효과의 동질성 여부 그리고 성향 점수 분포의 세 가지 중복(distribution overlap) 수준 하에서 실제 평균 치료 효과를 얼마나 강건하게 추정하는지 알아본다. 나아가 유한 표본에서 성향 점수 모델의 잘못된 지정과 보류된 치료가 존재하는 경우 각 가중치별 성능을 평가하기 위한 모의실험 연구를 시행하였다. 위의 상황에서 PS 모델이 올바르게 지정된 경우 OW와 MW는 IPW와 절사IPW보다 더 강건한(robust) 치료 효과를 추정한다는 것은 알 수 있다 (Crump 등, 2009). 그러나 PS 모델이 잘못 지정되었을 때와 예측에 반하는 치료가 발생한 경우의 각 가중치별 성능에 대해서는 알려진 바가 거의 없다. 우리는 모의실험을 위해 1,000명의 환자 정보를 이용하여 각 상황마다 1,000번의 몬테카를로 시뮬레이션을 시행하여 평가하였다.

5.2. 모의실험 설계

고려된 각 가중치 방법은 서로 다른 관심 효과 추정치를 대상으로 한다. 따라서 모의실험 전반에 걸쳐 참(true) 계수, 공변량 및 모델을 기반으로 크기 10^7 단위의 "초모집단(superpopulation)"을 생성하여 이진 처리 효과에서 IPW, trimmed IPW($\alpha = 0.05, 0.10, 0.15$), OW, MW 에 대한 치료 효과 참값을 계산했다.

우리는 Li와 Greene (2013)의 데이터 생성 과정을 기반으로 공변량 $\boldsymbol{X} = (X_1, \ldots, X_6)$ 와 두 가지 치료군 Z 를 생성했다. 이에 따라 $X_4 \sim \text{Ber}(0.5)$, $X_3 \sim \text{Ber}(0.4 + 0.2X_4)$, $X_5 = X_1^2$, $X_6 = X_2X_4$,

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} X_4 - X_3 + 0.5X_3X_4 \\ -X_4 + X_3 + X_3X_4 \end{pmatrix}, \begin{pmatrix} 2 - X_3 & 0.25(1 + X_3) \\ 0.25(1 + X_3) & 2 - X_3 \end{pmatrix}\right),$$

$Z \sim \text{Bernoulli}(e(\boldsymbol{X}))$, $e(\boldsymbol{X}) = [1+\exp\{-(\beta_0+\beta_1X_1+\cdots+\beta_6X_6)\}]^{-1}$ 을 만족한다. 이 때, 치료군과 대조군의 비율이 비슷한 경우 성향 점수 분포는 중복 수준이 좋음(good overlap), 적당함(moderate overlap), 좋지 않음(poor overlap)에 따라 $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6) = (-0.5, 0.3, 0.4, 0.4, 0.4, -0.1, -0.1)$, $(-1, 0.6, 0.8, 0.8, 0.8, -0.2, -0.2)$, $(-1.5, 0.9, 1.2, 1.2, 1.2, -0.3, -0.3)$ 으로 지정한다. 전체 피험자 중 치료군의 비율이 낮은(low) 경우 β_0를 각 중복 수준에 따라 $-1.5, -3, -4.5$ 로 지정하여 위와 동일한 과정의 분포를 만들었다. 그리고 우리는 치료 효과가 동질적(homogeneous)인 상황의 $\Delta = 1$, 이질적(heterogeneous)인 상황의 $\Delta = -4e(\boldsymbol{X})^2 + 3.94e(\boldsymbol{X}) + 0.69$ 으로 설정하여 최종 치료 결과 $Y \sim 0.5 + \Delta Z + X_1 + 0.6X_2 + 2.2X_3 + 1.2X_4 + 0.1X_5 + X_6 + N(0, 1)$ 으로 얻을 수 있다.

본 실험의 환자 집단 설계에서 치료 효과가 동질적일 때는 개인별로 편차가 존재하지 않고 모든 환자에서 치료 효과가 동일하게 나타나고, 이질적일 때는 환자의 특성에 따라 치료에 차이가 두드러지게 나타난다. 치료군과 대조군의 비율이 비슷한 경우를 중간 치료 확률(medium prevalence), 전체 피험자 중 치료를 받을 환자의 비율이 낮은 경우를 낮은 치료 확률(low prevalence)이라고 한다. 치료군과 대조군의 치료 할당 확률 분포 공통영역이 큰 경우를 좋은 중복, 작은 경우를 좋지 않은 중복, 좋은 중복과 좋지 않은 중복 사이를 적당한 중복이라 한다. 중간 치료 확률에서 성향 점수 분포의 중복 수준에 따른 변화는 Figure 1에 나타나있다. 해당 모의 실험은 McDonald 등 (2013)의 PS범위 중복(range overlap)과는 다르게 Mao와 Li (2020), Arisido 등 (2022)이 고안해낸 PS분포 중복 정도에 따라 성능을 비교했다.

나아가 실제 임상 상황에서 발생 할 수 있는 두 가지 문제 상황을 모의실험에 반영하였다. 첫 번째는 성향 점수 분포의 꼬리에 있는 환자에게 예측에 반하는 치료가 종종 발생하게 되는 문제 상황이다. 고위험(frailty)의 환자 중 일부가 치료 담당의에 의해 너무 허약하다고 간주되어서 자신의 임상 상태에 따라 받을 수 있는 치료를 받지 못하게 된 경우이다 (Stürmer 등, 2010; Zhang 등, 2019). 이러한 상황을 모의실험에 상정하기 위해 성향 점수가 상위 10%임과 동시에 치료 그룹에 할당된 대상 중에서 그 수를 무작위으로 선택하여 대조그룹으로 임의로 할당한 뒤 그에 따른 평균 치료 효과를 추정하였다. 두 번째는 성향 점수를 추정하는 과정에 측정되지 않은 교란요인이 존재하여 성향 점수 모형을 잘못 지정되게 하여 발생하는 상황이다. 측정되지 않은 교란요인에 의해 성향 점수 모형이 잘못 지정된 경우를 우리는 공변량 X_2, X_6을 하나씩 생략(omission)하여 확인하였다.

Zhou 등 (2020)과 Austin (2022)의 연구를 참고하여 이상의 과정을 1,000회 반복하여 문제 상황별, 가중치 방법별 실제 치료 효과와 추정된 치료 효과 사이의 편향(bias), 제곱근평균제곱오차(root mean square error; RMSE), 경험적 표준편차(standard deviation; SD), 평균 표준오차(standard error; SE), 포함 확률(coverage probability; CP)을 기준으로 성능을 평가하였으며, 실제 치료 효과는 고려된 데이터 생성 과정을 기반으로 초모집단에서 제공된 참(true) 추정값을 나타낸다. 모든 통계 분석은 R 4.1.0 버전을 이용하여 수행하였다.

5.3. 모의실험 결과

중간 치료 확률이며 환자 간의 치료 효과가 동질적일 때 모의실험 결과를 Table 2에 정리하였다. 각 표에서 complete, withheld, missing X_2, missing X_6는 네 가지 실험 조건을 각각 나타낸다. 표에서 complete는 치료 효과 추정 과정에서 어떠한 문제도 발생하지 않고 완전하게 이뤄진 상황이며, withheld는 성향 점수 모델 추정은 올바르게 되었으나 고위험군 환자에 의해 예측에 반하는 치료의 보류가 나타난 상황을 의미한다. 성향 점수 모델 추정 과정에서 공변량을 생략한 두 가지 문제 상황을 missing X_2, missing X_6로 Table 2의 하단에 같이 나타냈다. 위 네 가지 실험 조건하에서 치료군 간의 중복 상황이 변함에 따라 동일하게 시행된 다양한 가중치 방법 성능의 결과를 표에 보고하였다. 또한 이에 해당하는 공변량의 절대 표준화 평균 차이(absolute standard mean difference; ASD) 그림은 Figure 2에 주어져 있다.

IPW는 예측에 반하는 치료가 나타나거나 PS 모형이 잘못 추정되는 문제가 발생한 상황 뿐만 아니라 PS의 분포가 좋지 않은 중복 상황에서 더 민감하다는 것을 Table 2에서 알 수 있다. 평균 치료 효과를 추정하는데 문제가 없는 complete 시나리오의 좋은 중복 상황에서는 IPW를 포함한 다른 가중치 방법들 또한 작은 편향과 작은 분산으로 인해 믿을만한 추정치를 제공한다. 그럼에도 불구하고 중복 상황이 나빠지게 되면 IPW의 편향과 표준오차의 값이 다른 가중치 방법들에 비해 급격하게 증가하는 경향을 보인다. 또한 PS 모형이 잘못 지정된 상황에서는 IPW를 이용한 치료 효과 추정값이 부정확해 지는데, 이는 중복의 정도가 낮아질수록 더 심해지는 경향을 보인다. 위와 같은 경향은 예측에 반하는 치료가 존재하는 경우인 withheld 상황에서도 확인할 수 있는데 중복 정도가 낮아짐에 따라 IPW의 성능이 다른 방법들에 비해 가장 악화되는 것이 나타난다. 표의 상대적 편향이 10^2을 곱한 척도로 제시되었다는 사실을 고려할 때 IPW추정값과 다른 가중치 방법 간의 편향 차이에 비해 표준오차의 차이는 간과할 수 없는 차이라고 생각할 수 있다.

앞선 상황에서 IPW와는 다르게 임상적 균형을 이루는 것을 목표로 하는 가중치 OW와 MW는 X_2와 X_6이 생략된 경우에도 치료 효과를 추정하는데 있어 편향과 분산 측면에서 상대적으로 더 나은 결과값을 가지는 것을 알 수 있다. 특히 OW는 예측에 반하는 치료가 존재하는 상황에서는 다른 방법들에 비해 성능이 월등하게 좋은 것으로 확인된다. 이 방법은 두 집단간 중복되는 특성을 갖고 있는 환자에게는 큰 가중치를 주는 방식을 통해 그들을 강조하며, 성향 점수가 극단적인 환자들은 작은 가중치를 부여하는 특성을 갖고 있기 때문에 예측에 반하는 치료가 발생한 상황에서 OW가 상대적으로 더 좋은 결과를 보여주는 것이라 추측할 수 있다. 절사IPW의 편향과 분산 값이 IPW보다 상향된 성능을 보이긴 하지만, 대체적으로 OW와 MW에 비해서는 효율성이 감소하거나 부정확한 추정치를 보이는 것을 확인할 수 있다. 좀 더 효율적인 절사IPW를 위해 적절한 임계값을 고려해 보았을 때, $\alpha = 0.15$으로 설정한 절사 방법이 가장 정확한 치료 효과 추정값을 이끌어내기에 적합하다는 것이 나타난다. 이를 바탕으로 IPW보다는 OW와 MW가 PS 모형이 잘못 지정된 상황 또는 예측에 반하는 치료가 발생한 상황에서도 좀 더 효율적으로 추정치에 가까운 값을 찾아냄을 알 수 있다.

주요 변수인 X_2가 생략되었을 때의 결과를 Table 2의 하단부에서 확인할 수 있는데, 앞서 논의한 예측에 반하는 치료상황을 가정했을 때의 결과와 상이함을 볼 수 있다. X_2가 생략된 경우에는 어떠한 가중치 방법도 만족할 만한 결과를 도출하지 못하였는데, 특히 이 경우 95% 포함 확률은 성능을 비교하기에 정확한 정보를 제공하지 못하는 것으로 나타난다. IPW의 95% 포함확률이 동일한 조건하의 다른 가중치 방법보다 더 넓은 범위를 포함하는 것은 편향과 팽창된(inflated) 표준오차의 크기를 고려할 때 다소 부정확하게 과장된 신뢰구간으로 해석될 수 있다. 따라서 X_2가 치료와 결과의 인과관계에 큰 영향력을 주는 공변량이므로 이 변수를 PS 모형에서 제외시킬경우 결과의 신뢰도가 현저히 떨어짐을 확인할 수 있다.

Figure 2는 각 가중치 방법에 따른 모든 공변량의 ASD값을 그림으로 나타내었다. 가중치를 주지 않았을 때(unweighted)의 ASD값이 0.1보다 매우 큰 것으로 보아, 모의실험 자료의 치료 그룹간 공변량이 불균형하다는 것을 알 수 있다. 중복 상황에 상관없이 성향 점수 가중치를 통해 조정된(adjusted) ASD값은 대체적으로 0.1보다 작거나 근접한 값을 가지는 것을 확인할 수 있다. 또한, 예측에 반하는 치료가 일어난 경우와 다양한

PS 중복조건의 상황에서도 마찬가지로 모든 가중치 방법이 공변량의 균형을 잘 맞추는 것을 알 수 있다. 다만 좋지 않은 중복 상황의 경우 IPW의 ASD값이 0.1을 미미하게 벗어나는 경우를 볼 수 있다. 마지막으로 OW는 시나리오 전반에서 항상 정확한 평균 균형을 이끌어내는것을 확인할 수 있다.

추가로, Table 3에는 중간 치료 확률에서 치료 효과가 이질적일 때 수행한 모의실험의 결과를 나타낸다. 이질적 치료상황 하에서 추정량의 실제값의 범위는 좋은 중복에서는 1.53–1.57, 적당한 중복에서는 1.33–1.48, 좋지 않은 중복에서는 1.17–1.47로 계산되었고, 이때의 성능은 Table2에 나타나 있는 동질적인 치료 상황에서의 값과 비슷하다. 즉, IPW는 개인별 치료 효과의 동질성 여부보다는 성향 점수 분포의 겹침 정도에 더욱 민감하게 반응한다. PS 분포의 중복 수준이 낮아짐에 따라 0과 1에 가까운 극단적인 성향 점수를 갖는 피험자의 수가 많아지게 되는데 이러한 피험자의 IPW 값이 비정상적으로 매우 크거나 작은 경우가 다수 발생하게 된다. 이러한 이유에 따라 좋지 않은 중복 상황에서 과대 또는 과소 추정된 IPW를 이용하여 얻은 치료 효과는 실제 치료효과를 정확하게 추정하지 못하는 것으로 추론된다. 전반적으로 절사IPW보다 OW와 MW가 대체로 더 우수하게 결과를 보였으며, 예측에 반하는 치료가 있는 경우에는 OW의 성능이 가장 좋은 결과를 가지는 것을 알 수 있다. 또한 중요한 교란 변수 X_2가 분석에서 누락된 경우는 실질적으로 고려한 모든 가중치 방법이 제대로 된 추정치를 계산해내지 못하는 것을 확인하였다.

Tables 4–5에는 낮은 치료 확률에서 치료 효과가 동질적일 때와 이질적일 때 모의실험 결과를 나타낸다. 중간 치료 확률의 좋은 중복, 적당한 중복, 좋지 않은 중복에서 치료를 받을 환자 비율은 각 47.35%, 46.58%, 46.41%에 해당하며 치료군과 대조군의 비율이 비슷하다. 낮은 치료 확률에서는 중복 정도별 각 27.17%, 17.27%, 17.41%의 피험자가 치료를 받을 것으로 기대되어 전체 피험자 중 치료군에 배정된 환자의 수가 매우 불균형하게 적은 것을 확인할 수 있다. 시험 결과 낮은 치료 확률에서의 성능은 앞의 중간 치료 확률 상황에서의 흐름과 비슷하다. 특히 IPW의 경우 낮은 치료 확률에서 중복 수준이 낮아짐에 따라 성능이 더욱 급격하게 나빠지는 것을 볼 수 있다. 전반적으로 PS 분포의 중복 수준이 가중치 방법별 성능을 판단하는 데 가장 중요한 요소임을 모의실험 전체의 상황에서 확인하였다. Tables 2–5의 모의실험 결과에 따르면 절사IPW, OW, MW 세 가지 방법은 IPW의 단점을 개선하였으며, 대체적으로 OW가 다양한 상황에서 좋은 성능을 보이는 것을 확인할 수 있다.

6. 결론

본 논문에서는 평균 치료 효과를 추정할 때 교란요인을 보정하기 위하여 사용하는 다양한 성향 점수 가중치 방법을 간단하게 살펴보고, 모의실험을 통하여 비교하였다. 비교실험 결과 IPW에 비해 절사IPW의 결과값이 좀 더 치료 효과를 잘 추정하는 것을 볼 수 있었는데, 그 결과는 절사 임계값에 따라 차이가 있었다. 따라서 가장 적정한 추정치를 제공하는 임계값을 결정하는 것은 많은 주의가 필요하다고 볼 수 있으며 임상적 배경지식에 기반한 임계값 결정 과정이 필요할 것으로 생각된다. 그 예로 Glynn 등 (2019)은 유병률 정보를 이용한 임계값 α를 0.1로 이용하여 아토르바스타틴(atorvastatin)과 로수바스타틴(rosuvastatin)사이의 치료 효과를 비교했다. 또한 OW와 MW를 이용한 치료 효과의 추정값이 대체적으로 IPW와 절사 IPW에 비해 더 안정적인 결과를 보였다. 따라서 위의 두 방법이 IPW 기반의 방법에 비해 좀 더 강건(robust)한 추정치를 제공한다는 것을 알 수 있으며, 그중에서도 OW의 성능이 가장 우수하기에 널리 활용 가능한 방법이라 할 수 있다.

본 연구에서는 예상치 못하게 치료를 받지 못한 고위험군 환자를 성향 점수 분포 상위 10% 내에서 무작위로 선택하여 실험을 진행하였는데, 이보다 더 임상 상황을 실제적으로 반영하는 실험 설계를 구현하여 가중치를 비교한다면 좋은 연구가 될 수 있을 것이다. 또한 본 논문에서는 PS 분포 중복 정도에 따른 가중치의 성능을 비교했는데, 여기서 더 나아가 PS 범위 중복 정도에 따른 가중치 성능을 비교하는 것도 좋은 후속 연구가 될 것이라 생각된다.

References

Arisido MW, Mecatti F, and Rebora P (2022). Improving the causal treatment effect estimation with propensity scores by the bootstrap, *AStA Advances in Statistical Analysis*, **106**, 455–471.

Austin PC (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003, *Statistics in Medicine*, **27**, 2037–2049.

Austin PC (2022). Bootstrap vs asymptotic variance estimation when using propensity score weighting with continuous and binary outcomes, *Statistics in Medicine*, **41**, 4426–4443.

Austin PC and Stuart EA (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies, *Statistics in Medicine*, **34**, 3661–3679.

Cochran WG and Rubin DB (1973). Controlling bias in observational studies: A review, *Sankhyā: The Indian Journal of Statistics, Series A*, **35**, 417–446.

Crump RK, Hotz VJ, Imbens GW, and Mitnik OA (2009). Dealing with limited overlap in estimation of average treatment effects, *Biometrika*, **96**, 187–199.

Freedman DA and Berk RA (2008). Weighting regressions by propensity scores, *Evaluation Review*, **32**, 392–409.

Glynn RJ, Lunt M, Rothman KJ, Poole C, Schneeweiss S, and Stürmer T (2019). Comparison of alternative approaches to trim subjects in the tails of the propensity score distribution, *Pharmacoepidemiology and Drug Safety*, **28**, 1290–1298.

Godambe VP (1970). Foundations of survey-sampling, *The American Statistician*, **24**, 33–38.

Hirano K, Imbens GW, and Ridder G (2003). Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, **71**, 1161–1189.

Joffe MM and Rosenbaum PR (1999). Invited commentary: Propensity scores, *American Journal of Epidemiology*, **150**, 327–333.

Kim B and Kim JH (2020). Estimating causal effect of multi-valued treatment from observational survival data, *Communications for Statistical Applications and Methods*, **27**, 675–688.

Kim GS, Paik MC, and Kim H (2017). Causal inference with observational data under cluster-specific non-ignorable assignment mechanism, *Computational Statistics & Data Analysis*, **113**, 88–99.

Lee BK, Lessler J, and Stuart EA (2011). Weight trimming and propensity score weighting, *PloS One*, **6**, 1–6.

Li F, Morgan KL, and Zaslavsky AM (2018). Balancing covariates via propensity score weighting, *Journal of the American Statistical Association*, **113**, 390–400.

Li F, Thomas LE, and Li F (2019). Addressing extreme propensity scores via the overlap weights, *American Journal of Epidemiology*, **188**, 250–257.

Li L and Greene T (2013). A weighting analogue to pair matching in propensity score analysis, *The International Journal of Biostatistics*, **9**, 215–234.

Lunceford JK and Davidian M (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study, *Statistics in Medicine*, **23**, 2937–2960.

Mao H and Li L (2020). Flexible regression approach to propensity score analysis and its relationship with matching and weighting, *Statistics in Medicine*, **39**, 2017–2034.

Mao H, Li L, and Greene T (2019). Propensity score weighting analysis and treatment effect discovery, *Statistical Methods in Medical Research*, **28**, 2439–2454.

McDonald RJ, McDonald JS, Kallmes DF, and Carter RE (2013). Behind the numbers: Propensity score analysis—a primer for the diagnostic radiologist, *Radiology*, **269**, 640–645.

Robins J (1986). A new approach to causal inference in mortality studies with a sustained exposure period—appli cation to control of the healthy worker survivor effect, *Mathematical Modelling*, **7**, 1393–1512.

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rosenbaum PR and Rubin DB (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, *The American Statistician*, **39**, 33–38.

Rubin DB (1973). Matching to remove bias in observational studies, *Biometrics*, **29**, 159–183.

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, **66**, 688–701.

Rubin DB (1980). Randomization analysis of experimental data: The fisher randomization test comment, *Journal of the American Statistical Association*, **75**, 591–593.

Stefanski LA and Boos DD (2002). The calculus of *m*-estimation, *The American Statistician*, **56**, 29–38.

Stuart EA (2010). Matching methods for causal inference: A review and a look forward, *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, **25**, 1–21.

Stürmer T, Rothman KJ, Avorn J, and Glynn RJ (2010). Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution—a simulation study, *American Journal of Epidemiology*, **172**, 843–854.

Stürmer T, Webster-Clark M, Lund JL, Wyss R, Ellis AR, Lunt M, Rothman KJ, and Glynn RJ (2021). Propensity score weighting and trimming strategies for reducing variance and bias of treatment effect estimates: A simulation study, *American Journal of Epidemiology*, **190**, 1659–1670.

Traskin M and Small DS (2011). Defining the study population for an observational study to ensure sufficient overlap: A tree approach, *Statistics in Biosciences*, **3**, 94–118.

Zhang HT, McGrath LJ, Ellis AR, Wyss R, Lund JL, and Stürmer T (2019). Restriction of pharmacoepidemiologic cohorts to initiators of medications in unrelated preventive drug classes to reduce confounding by frailty in older adults, *American Journal of Epidemiology*, **188**, 1371–1382.

Zhou Y, Matsouaka RA, and Thomas L (2020). Propensity score weighting under limited overlap and model misspecification, *Statistical Methods in Medical Research*, **29**, 3721–3756.

Received March 12, 2023; Revised May 4, 2023; Accepted May 8, 2023

임상에서 발생할 수 있는 문제 상황에서의 성향 점수 가중치 방법에 대한 비교 모의실험 연구

정시성[a], 민은정[1,a,b]

[a]가톨릭대학교 의생명 · 건강과학과; [b]가톨릭대학교 의과대학 의생명과학교실

요 약

대부분의 임상시험에서 가장 대표적으로 사용되는 실험설계는 무작위화로, 치료 효과를 정확하게 추정하기 위해 이용된다. 그러나 무작위화가 이루어지지 않은 관찰연구의 경우 치료군과 대조군의 비교로 얻는 치료 효과에는 환자 간의 특성 등 여러 조정되지 않은 차이로 인해 편향이 발생한다. 성향 점수 가중치는 이러한 문제점을 해결하기 위해 널리쓰이는 방법으로 치료 효과를 추정하는데에 있어 교란요인을 조정하여 편향을 최소화하도록 하는 방법이다. 성향 점수를 이용한 가중치 방법 중 가장 널리 알려진 역확률 가중치는 관찰된 공변량이 주어졌을 때 특정 치료에 할당될 조건부 확률의 역에 비례하는 가중치를 할당한다. 그러나 이 방법은 극단적인 성향 점수에 의해 종종 방해 받아 편향된 추정치와 과도한 분산을 초래한다는 점이 알려져있어 이러한 문제를 완화하기 위해 절사 역확률 가중치, 중복 가중치, 일치 가중치를 포함한 여러 가지 대안 방법이 제안되었다. 본 논문에서는 제한된 중복, 잘못 지정된 성향 점수 모델 및 예측과 반대되는 치료 등 다양한 문제 상황에서 여러 성향 점수 가중치 방법의 성능을 비교하는 시뮬레이션 비교연구를 수행하였다. 비교연구의 결과 중복 가중치와 일치 가중치는 편향, 제곱근평균제곱오차 및 포함 확률 측면에서 역확률 가중치와 절사 역확률 가중치에 비에 우월한 성능을 보임을 확인하였다.

주요용어: 성향 점수, 역확률 가중치, 모의실험, 제한된 중복

이 논문은 2021년도 정부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2021R1F1A1058613).

[1]교신저자: (06591) 서울특별시 서초구 반포대로 222, 가톨릭대학교 의과대학. E-mail: ej.min@catholic.ac.kr