傾向スコアの概念とその実践

奥村泰之
一般財団法人 医療経済研究・社会保険福祉協会
医療経済研究機構 研究部 研究員

第4回臨床研究実践講座ワークショップ
2015/1/23 (金) 13:45~14:30
国立精神・神経医療研究センター
研究所3号館セミナールーム

構成

- ■特徴
- ■共変量の選択
- ■傾向スコアの推定
- ■傾向スコアの利用
- ■バランスの評価
- ■効果の推定
- ■効果の解釈
- ■傾向スコアの実践

英語表記

- Propensity (Score) Analysis/Methods/Matching
- Matching Methods

変数の役割と尺度水準

■1つのアウトカム

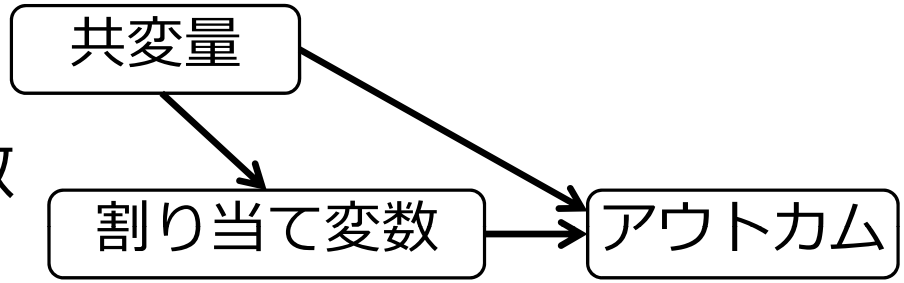
- ●量的変数/質的変数/イベント発生までの時間
  - 生活の質，生きている/死んでいる，生存時間

■1つの割り当て変数

- ●名義尺度 (2水準が中心)
  - 曝露群/非曝露群

■1つ以上の共変量

- ●量的変数/質的変数



4

傾向スコア分析の使用目的

- アウトカムの測定後に傾向スコア分析を行い，選択バイアスを減らして，治療の効果を検討する
- アウトカムの測定前に傾向スコア分析を行い，追跡する集団を限定する

Stuart EA: Stat Sci 25:1-21, 2010 .

事例①測定後

■目的
- ◆認知症への非定型抗精神病薬と定型薬の死亡リスク

■アウトカム
- ◆抗精神病薬の投与開始後180日以内の全死亡

■割り当て変数
- ◆非定型薬 vs 定型薬

■共変量
- ◆認知症重症度，90日前の入院の有無，年齢など

Comparison	Atypical Antipsychotic Use	Conventional Antipsychotic Use	Absolute Risk Difference (95% CI), *percentage points**
Conventional vs. atypical antipsychotic use at 180 d			
Community-dwelling cohort	10.7	13.3	2.6 (0.5 to 4.5)
Long-term care cohort	17.8	20.0	2.2 (0.0 to 4.4)

Gill SS et al: Ann Intern Med 146:775-86, 2007.

6

事例②測定前

■目的
- ◆妊娠中のフェノバルビタールによる胎児成長後の知能

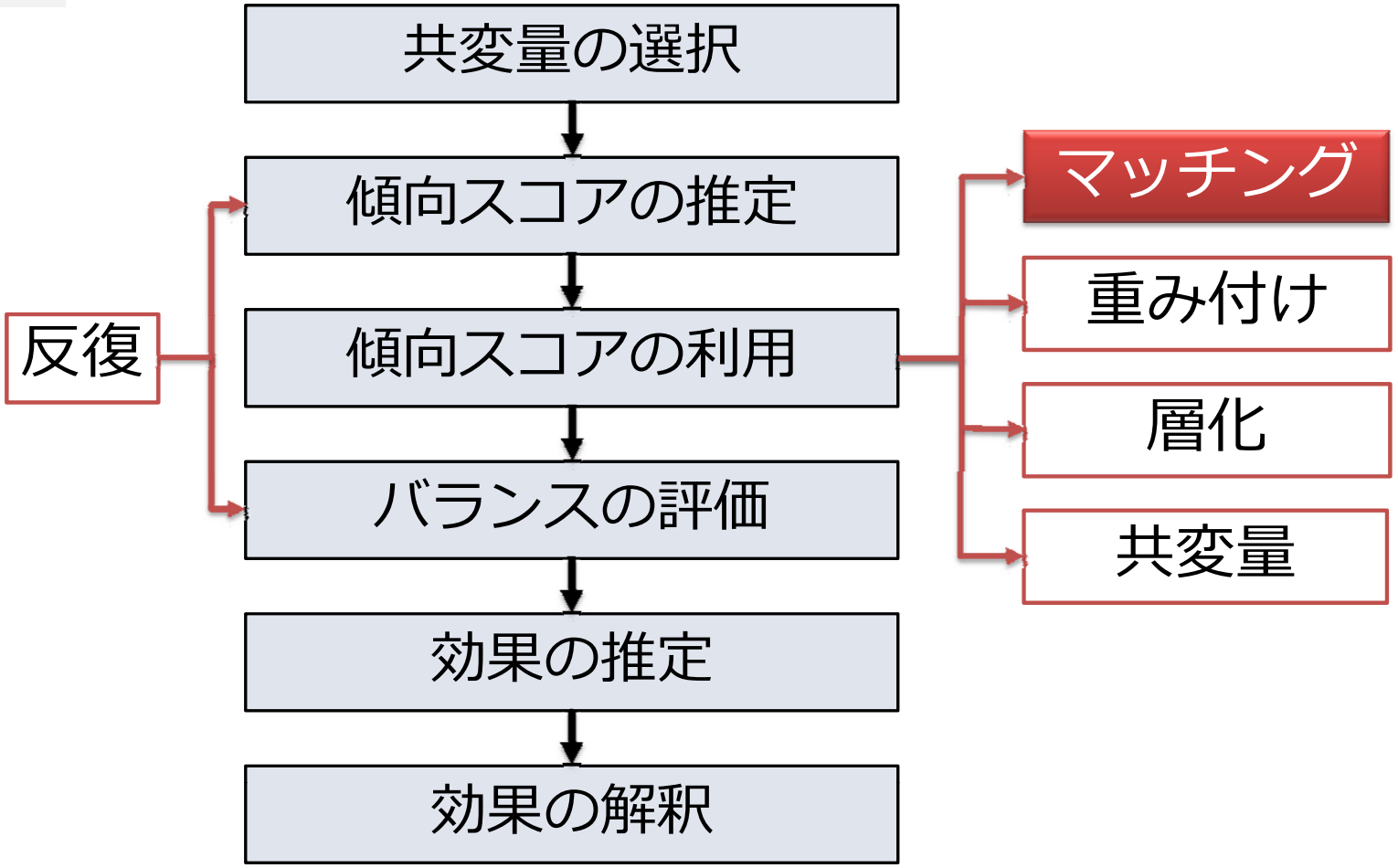
■アウトカム
- ◆胎児の成長後の知能

■割り当て変数
- ◆処方あり (33名) vs 処方なし (3308名)

■共変量
- ◆社会経済的地位，父親の有無など

Reinisch JM et al: JAMA 274: 1518-25, 1995.

解析の手順



Ali MS et al: J Clin Epidemiol. 2014 Nov 26. pii: S0895-4356(14)00347-3

報告ガイドライン (赤字は推薦文献)

- Ali MS et al: Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review (J Clin Epidemiol. 2014)
- D'Ascenzo F et al: Use and misuse of multivariable approaches in interventional cardiology studies on drug-eluting stents: a systematic review (J Interv Cardiol. 2012 Dec;25(6):611-21.)
- Collins GS et al: Comparing treatment effects between propensity scores and randomized controlled trials: improving conduct and reportin (Eur Heart J 33:1867-9, 2012)
- Gayat E et al: Propensity scores in intensive care and anaesthesiology literature: a systematic review (Intensive Care Med 36: 1993-2003, 2010)
- Austin PC: A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003 (Stat Med 27:2037-49, 2008)

報告ガイドライン (赤字は推薦文献)

- Austin PC: Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review (Circ Cardiovasc Qual Outcomes 1: 62-7, 2008)
- Austin PC: Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement (J Thorac Cardiovasc Surg 134:1128-35, 2007)
- Stürmer T et al: A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods (J Clin Epidemiol 59: 437-47, 2006)
- Shah BR et al: Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review (J Clin Epidemiol 58: 550-9, 2005)
- Weitzen S et al: Principles for modeling propensity scores in medical research: a systematic literature review (Pharmacoepidemiol Drug Saf 13: 841-53, 2004)

共変量の選択

共変量の適格基準

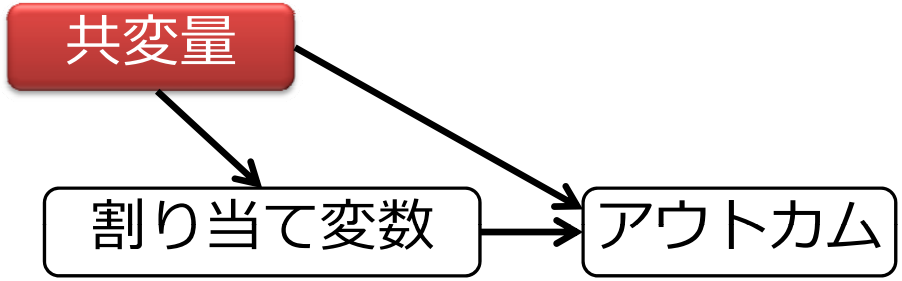
■選択基準

◆割り当て変数の前/同時に測定された変数

■除外基準

◆アウトカム

◆中間変数 (割り当て変数により変化しうる変数)

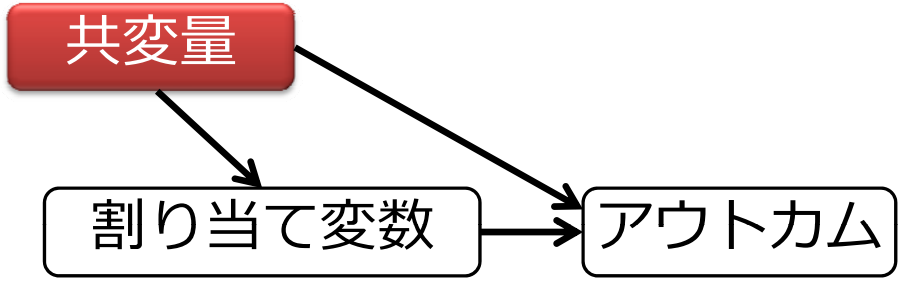


Austin PC et al: Analysis of observational health care data using SAS (pp51-84). SAS press. 2010.

12

共変量の選択法

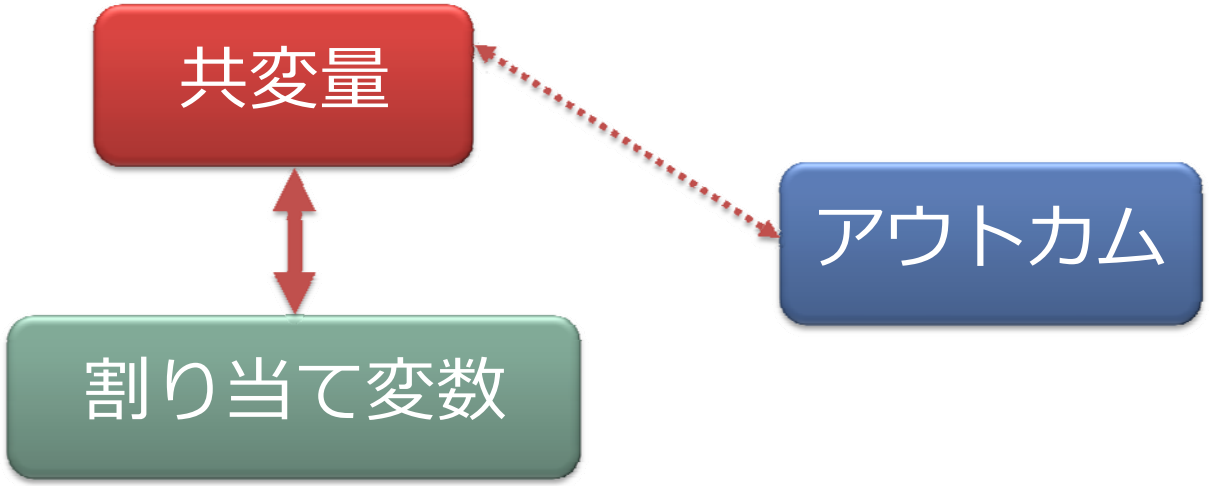
- 先行研究
- 専門家の意見
- 統計量 (ステップワイズ法など)



Austin PC: Stat Med 27:2037-49, 2008.

避けるべき共変量

■「割り当て変数と強く関連」かつ「アウトカムと弱く関連」



Brookhart MA et al: Am J Epidemiol. 2006 Jun 15;163(12):1149-56.
Patrick AR et al: Pharmacoepidemiol Drug Saf. 2011 Jun;20(6):551-9.

事例①先行研究

- The PS model was estimated using a logistic regression model that adjusted for the patient characteristics listed in Table 1, as well as admissions vital signs and laboratory values (31 variables in total), as these variables were shown to be prognostically significant in other studies [18].

Patient characteristics[a]	$N = 5338$	With orders ($n = 4314$)	No orders ($n = 1024$)
Provider characteristic			
Hospital type			
Teaching ($n = 10$)	1420 (26.6)	1088 (25.2)	332 (32.4)
Community ($n = 60$)	3780 (70.8)	3117 (72.3)	663 (64.7)
Small ($n = 7$)	138 (2.6)	109 (2.5)	29 (2.8)
Demographic characteristics			
Age (years), mean ± SD	66.0 ± 13.7	65.6 ± 13.6	67.7 ± 14.2
Female	1757 (32.9)	1392 (32.3)	365 (35.6)

Abrahamyan L et al: Int J Qual Health Care 24:425-32, 2012.

事例②専門家の意見

- Possible confounders were chosen for their potential association with the outcome of interest based on clinical knowledge. The predicted probability of preprocedural stains was calculated by fitting a logistic regression model, using all clinically relevant variables as shown in Table 1.

	Preprocedural Statins, n (%)	No Preprocedural Statins, Mean (SD)
No. of patients	9104	3876
Demographics		
Mean age, y, SD	73.55±5.32	74.44±5.95
Male	6204 (68.1%)	2456 (63.4%)
Low income	1760 (19.3%)	721 (18.6%)

Ko DT et al: Circ Cardiovasc Qual Outcomes 4:459-66, 2011.

事例③統計量

- Covariates were carefully selected based on the assumption that none was affected directly by the intervention. Other a priori selected variables were planned for inclusion in the final statistical models. ...(中略)...
- Variables were considered for inclusion in the final models after calculating correlation coefficients, examining scatterplot matrices, and ensuring that the proportion of missing data was below 20%.

Galvagno SM et al: JAMA 307:1602-10, 2012.

傾向スコアの推定

傾向スコアとは

■正式な理解

◆観測した共変量が与えられた条件下で，ある因子に曝露する条件付き確率

■直感的な理解

◆ある人の曝露前の共変量を考慮したときに，ある人が曝露群になる確率

Trojano M et al: Int MS J 16: 90-7, 2009.

傾向スコア推定の統計モデル

- ロジスティック回帰分析 (logistic regression)
- プロビット回帰分析 (probit regression)
- 判別分析 (discriminant analysis)
- 決定木 (classification and regression trees)
- ニューラルネットワーク (neural networks)
- 一般化加法モデル (generalized additive models)
- 多項ロジットモデル (multinomial logistic regression)

Austin PC et al: Analysis of observational health care data using SAS (pp51-84). SAS press. 2010.

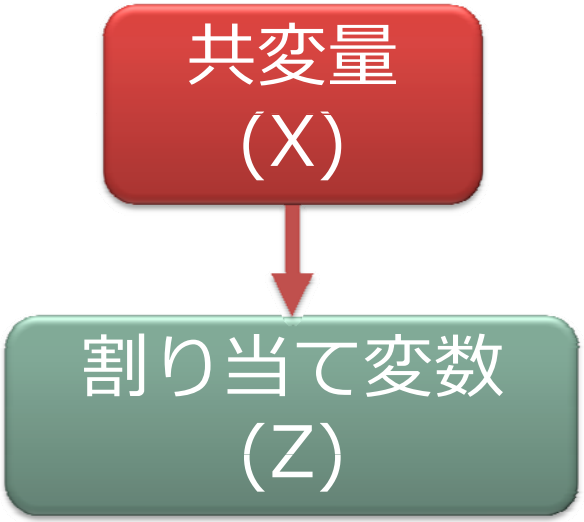
傾向スコア推定の変数役割

■従属変数
- ◆割り当て変数 (Z) (曝露群/非曝露群)

■独立変数
- ◆共変量 (X)

■傾向スコアの推定値 ($\hat{e}_i$)
- ◆予測値: $\hat{e}_i = pr(z_i = \text{曝露群} \mid x_i)$

■特徴
- ◆得点可能範囲: 0～1
- ◆サイズ: 標本サイズと同じ



21

事例: ロジスティック回帰分析

- Possible confounders were chosen for their potential association with the outcome of interest based on clinical knowledge. The predicted probability of preprocedural stains was calculated by fitting a logistic regression model, using all clinically relevant variables as shown in Table 1.

	Preprocedural Statins, n (%)	No Preprocedural Statins, Mean (SD)
No. of patients	9104	3876
Demographics		
Mean age, y, SD	73.55±5.32	74.44±5.95
Male	6204 (68.1%)	2456 (63.4%)
Low income	1760 (19.3%)	721 (18.6%)

Ko DT et al: Circ Cardiovasc Qual Outcomes 4:459-66, 2011.

傾向スコアの利用

傾向スコアの利用法

- マッチング (propensity score matching)
- 重み付け (inverse probability of treatment weighting)
- 層化 (stratification/subclassification)
- 共変量 (covariate adjustment)

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

マッチング法の設定

- ①アルゴリズム
- ②キャリパー
- ③構成比
- ④抽出法

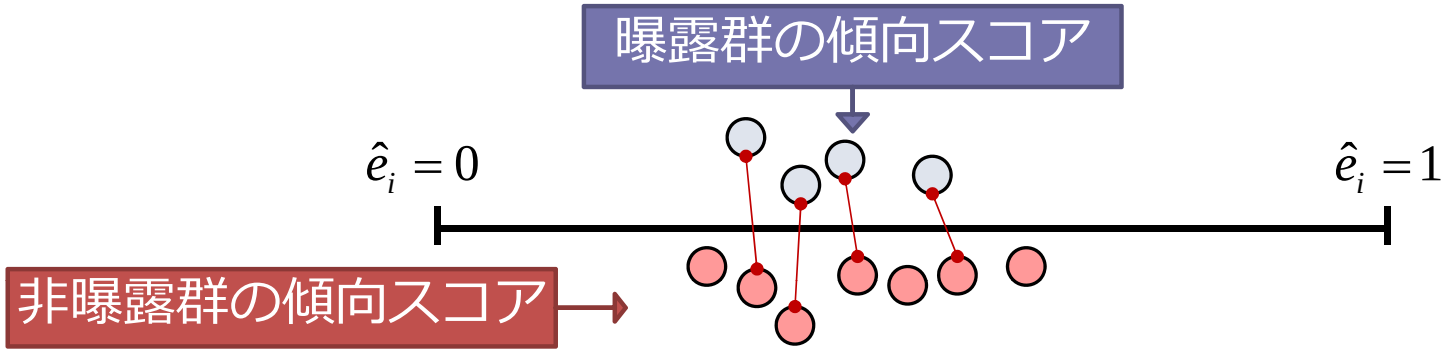
アルゴリズムの種類

- Greedy matching
  - Nearest neighbor matching
  - Mahalanobis metric matching
- Optimal matching

Guo S, Fraser MW: Propensity score analysis: statistical methods and applications. Sage. 2015.

Nearest neighbor matching

- ■①曝露群から無作為に1人選択
- ■②非曝露群から，①で選択した人の傾向スコアと，最も類似の傾向スコアの人をペアとする
- ■③上記の①~②を反復



Austin PC: Multivariate Behav Res 46: 399-424, 2011.

27

キャリパーの指定

- 指定法
  - 一定の傾向スコアの距離 (キャリパー) に収まる人をマッチングの対象とする
- 推奨値
  - 傾向スコアの推定値をロジット変換した値の標準偏差に0.2を乗じた値が推奨

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

傾向スコア間の距離を確認

ID	PS
t1	0.48
t2	0.97
t3	0.69
t4	0.68
t5	0.96
t6	0.34
c1	0.31
c2	0.00
c3	0.74
c4	0.02
c5	0.52
c6	0.29

曝露群		非曝露群		
ID	PS1	ID	PS2	|PS1-PS2|
t1	0.48	c5	0.52	0.04
t2	0.97	c3	0.74	0.23
t3	0.69	c1	0.31	0.37
t4	0.68	c6	0.29	0.39
t5	0.96	c4	0.02	0.93
t6	0.34	c2	0.00	0.34

一部
距離が大きい

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

29

一定距離内の人をマッチング

ID	PS
t1	0.48
t2	0.97
t3	0.69
t4	0.68
t5	0.96
t6	0.34
c1	0.31
c2	0.00
c3	0.74
c4	0.02
c5	0.52
c6	0.29

曝露群		非曝露群		
ID	PS1	ID	PS2	|PS1-PS2|
t1	0.48	c1	0.31	0.17
t2	0.97	c3	0.74	0.23
t3	0.69	c5	0.52	0.16
~~t4~~	~~0.68~~			
~~t5~~	~~0.96~~			
t6	0.34	c6	0.29	0.05

距離が0.3以上の曝露群を除外

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

30

構成比の種類

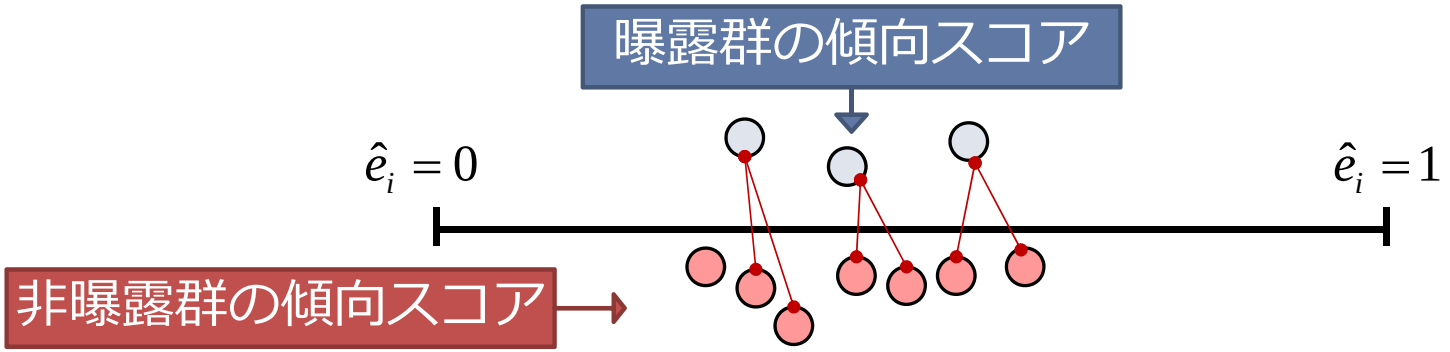
- One-to-one pair matching
  - 1名の曝露群と1名の非曝露群でマッチング
- Many-to-one (M:1) matching
  - 1名の曝露群とM名の非曝露群でマッチング
- Full matching
  - 1名の曝露群と複数名の非曝露群，1名の非曝露群と複数名の曝露群でマッチングの組み合わせ

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

31

Many-to-one (M:1) matching

■曝露群1人に，非曝露群を複数



32

抽出法の種類

■非復元抽出

◆曝露群のペアとして同一の非曝露群の人を，複数回使用できない

■復元抽出

◆曝露群のペアとして同一の非曝露群の人を，複数回使用できる

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

復元抽出

■同一の非曝露群を復元使用

ID	PS
t1	0.48
t2	0.97
t3	0.69
t4	0.68
t5	0.96
t6	0.34
c1	0.31
c2	0.00
c3	0.74
c4	0.02
c5	0.52
c6	0.29

曝露群		非曝露群		
ID	PS1	ID	PS2	|PS1-PS2|
t1	0.48	c5	0.52	0.04
t2	0.97	c3	0.74	0.23
t3	0.69	c3	0.74	0.06
t4	0.68	c3	0.74	0.07
t5	0.96	c3	0.74	0.21
t6	0.34	c1	0.31	0.03

非曝露群の
c3は4回使用

事例: nearest neighbor/0.2/4対1/非復元抽出

- Propensity matching was then performed according to both bleeding and mortality risk, using the nearest neighbor matching without replacement, with each bleeding patient matched to 4 control patients. A caliper width of 0.2 of the standard deviation of the logit of the propensity score was used for the developed propensity score, (後略)

Chhatriwalla AK et al: JAMA. 2013 Mar 13;309(10):1022-9.

バランスの評価

バランスの評価法

- 推奨法
  - 群ごとの要約統計量
  - 標準化差 (standardized difference)
  - QQ plotやside-by-side boxplot
- 非推奨法
  - 群間の共変量の差異の統計的検定
  - c統計量
  - 群ごとの傾向スコアの推定値の分布の比較

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

標準化差

■量的変数 $$d = \frac{(M_t - M_c)}{\sqrt{\frac{sd_t^2 + sd_c^2}{2}}}$$

■質的変数 $$d = \frac{p_t - p_c}{\sqrt{\frac{p_t(1-p_t) + p_c(1-p_c)}{2}}}$$

■特徴

◆得点可能範囲: −∞ ～ 0 (群間差なし) ～ +∞

◆0.1未満であればバランスが取れていると判断

◆*d*の絶対値に100を乗じる流派もある

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

バランスが悪いときの対処法

- 共変量を増やす
- 共変量間の交互作用項を追加
- 量的変数の共変量の非線形性を検討

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

事例: 要約統計量と標準化差

■方法の節

- We used a structured iterative approach to refine this logistic regression model to achieve balance of covariates within the matched pairs.[29] We used the standardised difference to measure covariate balance, whereby an absolute standardised difference above 10% represents meaningful imbalance.[29]

■結果の節

- Individuals who did and did not undergo consultation differed for all measured characteristics (Table 1).
- Of patients who underwent consultation, 91.6% (n=95 926) were matched to similar patients who did not. The covariate balance in the matched cohort was considerably improved (Table 2).

Wijeysundera DN et al: Arch Intern Med 170: 1365-74, 2010.

40

事例: 要約統計量と標準化差

標準化差

群ごとの要約統計量

マッチング後のバランス

Table 1. Characteristics of Individuals Who Did or Did Not Undergo Preoperati

Characteristic	No. (%)[a] Consultation (n=104 695)	No Consultation (n=165 171)	Absolute Standardized Difference, %
Demographics			
Age, mean (SD), y	70.1 (9.6)	67.3 (10.6)	27.3
Female sex	54 085 (51.7)	84 559 (51.2)	1.0
Neighborhood income, $			
Quintile 1 (highest)	22 166 (21.2)	33 412 (20.2)	2.5
Quintile 2	22 163 (21.2)	33 537 (20.3)	2.2
Quintile 3	20 660 (19.7)	32 038 (19.4)	0.8
Quintile 4	18 995 (18.1)	31 880 (19.3)	3.1
Quintile 5 (lowest)	20 185 (19.3)	33 248 (20.1)	2.0
Missing	526 (0.5)	1056 (0.6)	1.4

Table 2. Characteristics of the Propensity-Score Matched Pairs

Characteristic	No. (%)[a] Consultation (n=95 926)	No Consultation (n=95 926)	Absolute Standardized Difference, %
Demographics			
Age, mean (SD), y	69.8 (9.7)	69.9 (9.6)	0.4
Female sex	49 939 (52.1)	50 049 (52.2)	0.2
Neighborhood income, $			
Quintile 1 (highest)	20 192 (21.0)	20 101 (21.0)	0.2
Quintile 2	20 032 (20.9)	20 110 (21.0)	0.2
Quintile 3	18 909 (19.7)	18 895 (19.7)	0.04
Quintile 4	17 585 (18.3)	17 607 (18.4)	0.06
Quintile 5 (lowest)	18 707 (19.5)	18 699 (19.5)	0.02
Missing	501 (0.5)	514 (0.5)	0.2

Wijeysundera DN et al: Arch Intern Med 170: 1365-74, 2010.

41

効果の推定

効果推定の統計モデル

対応を考慮すべきか否か議論がある

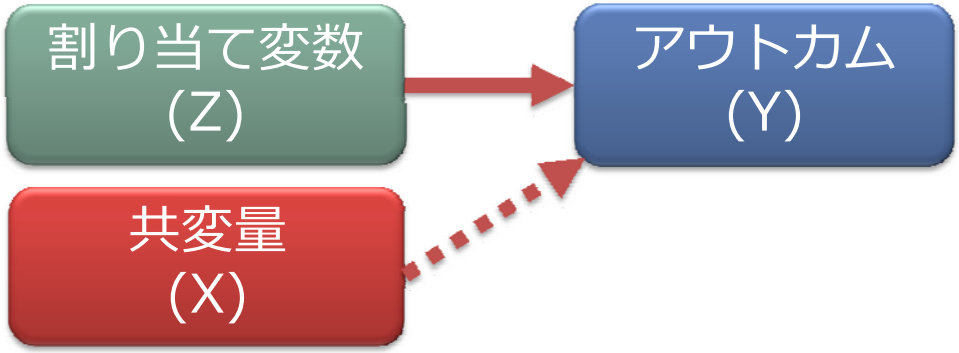
アウトカム	対応なし[1-2]	対応あり[3-4]
量的変数	独立な2群のt検定 U検定 一般化線形モデル	対応のある2群のt検定 ウィルコクソンの符号順位検定 一般化線形混合モデル 一般化推定方程式
質的変数	χ二乗検定 一般化線形モデル	マクネマー検定 AgrestiとMinの方法 条件付きロジスティック回帰分析 一般化線形混合モデル 一般化推定方程式
イベント発生までの時間	比例ハザードモデル	条件付き比例ハザードモデル ロバスト推定

[1] Stuart EA: Stat Med. 2008 May 30;27(12):2062-5 ; [2] Stuart EA: Stat Sci. 2010 Feb 1;25(1):1-21.
[3] Austin PC: Stat Med. 2008 May 30;27(12):2037-49.; [4] Austin PC: Stat Med. 2011 May 20;30(11):1292-301.

効果推定の変数役割

■従属変数
- ◆アウトカム

■独立変数
- ◆割り当て変数 (Z) (曝露群/非曝露群)
- ◆共変量 (X)



44

事例: 量的変数/質的変数/対応あり

■方法の節

- Within the matched pairs, we used the paired t test to compare hospital length of stay and the methods of Agresti and Min[28] to compare mortality rates.

■結果の節

- Within this matched cohort, mean hospital length of stay was significantly shorter among patients who underwent preoperative consultation (8.17 days vs 8.52 days; difference, −0.35 days; 95% confidence interval [CI], −0.27 to −0.43; P < .001). ...(中略)... Consultation was not associated with reduced mortality at either 30 days (relative risk [RR], 1.04; 95% CI, 0.96 to 1.13; P=.36) or 1 year (RR, 0.98; 95% CI, 0.95 to 1.02; P=.20) after surgery.

Wijeysundera DN et al: Arch Intern Med 169: 595-602, 2009.

効果の解釈

効果の種類

- 平均因果効果
- 曝露群の平均因果効果

Schafer JL, Kang J: Psychol Methods 13: 279-313, 2008.

平均因果効果 (Average Causal Effect)

■母集団の構成員すべてが曝露群から非曝露群に変化したときの，アウトカムの期待値の差

$$ACE = E\left(Y_{i\text{曝露}}\right) - E\left(Y_{i\text{非曝露}}\right)$$

ID	割り当て変数	アウトカム		
		曝露	非曝露	因果効果
1	非曝露	7.6	6.1	1.5
2	曝露	7.9	7.2	0.7
3	非曝露	4.1	5.2	-1.1
4	曝露	7.1	4.8	2.3
N	非曝露	8.3	6.9	1.4

注) 青色は欠測値

Schafer JL, Kang J: Psychol Methods 13: 279-313, 2008.

曝露群の平均因果効果 (Average Treatment effect for Treated)

■母集団の構成員のうち曝露群が非曝露群に変化したときの，アウトカムの期待値の差

$$ATT = E\left(Y_{i\text{曝露}} \middle| Z_i = \text{曝露}\right) - E\left(Y_{i\text{非曝露}} \middle| Z_i = \text{曝露}\right)$$

		アウトカム		
ID	割り当て変数	曝露	非曝露	因果効果
1	非曝露	7.6	6.1	
2	曝露	7.9	7.2	0.7
3	非曝露	4.1	5.2	
4	曝露	7.1	4.8	2.3
N	非曝露	8.3	6.9	

注) 青色は欠測値

Schafer JL, Kang J: Psychol Methods 13: 279-313, 2008.

曝露事例と効果の種類

曝露事例	効果の種類	理由
プライマリケアに受診する喫煙者へ禁煙を勧める冊子提供の効果	平均因果効果	喫煙者すべてに冊子を提供することは比較的安価
喫煙者への構造化された高強度の禁煙プログラムの効果	曝露群の平均因果効果	喫煙者すべてに禁煙プログラムを実施することは現実的でない

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

傾向スコアの利用法と効果の種類

利用法	平均因果効果	曝露群の平均因果効果
マッチング	×	○
重み付け	○	○
層化	○	○
共変量	○	○

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

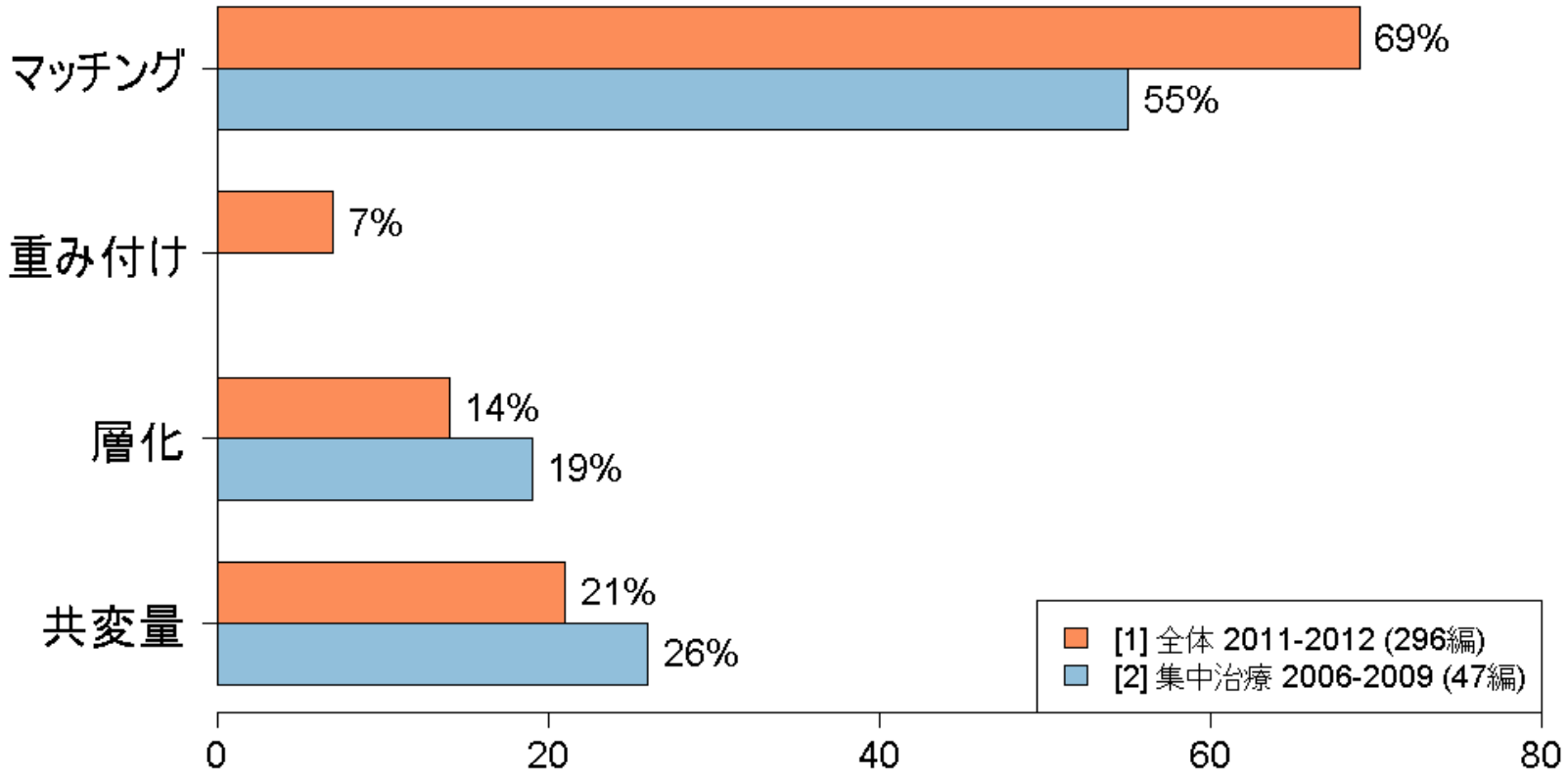
観測値からの因果効果推定の仮定

- ある人の割り当て変数により，他の人のアウトカムの値が変化しない
- 未測定の共変量はない
- 統計モデルが正しい
- ある人は曝露群か非曝露群のいずれかになり得る (片方の群となる確率が0や1でない)

Schafer JL, Kang J: Psychol Methods 13: 279-313, 2008.

傾向スコアの実践

傾向スコアの利用法



[1] Ali MS et al: J Clin Epidemiol. 2014 Nov 26. pii: S0895-4356(14)00347-3
[2] Gayat E et al: Intensive Care Med. 2010 Dec;36(12):1993-2003.

54

曝露群の標本サイズ

研究	第1四分位	中央値	第3四分位
[1] 集中治療 2006-2009 (47編)	121	968	1310
[2] 全体 1983-2003 (177編)	198	635	2247
[3] 全体 2001 (47編)	182	805	3802

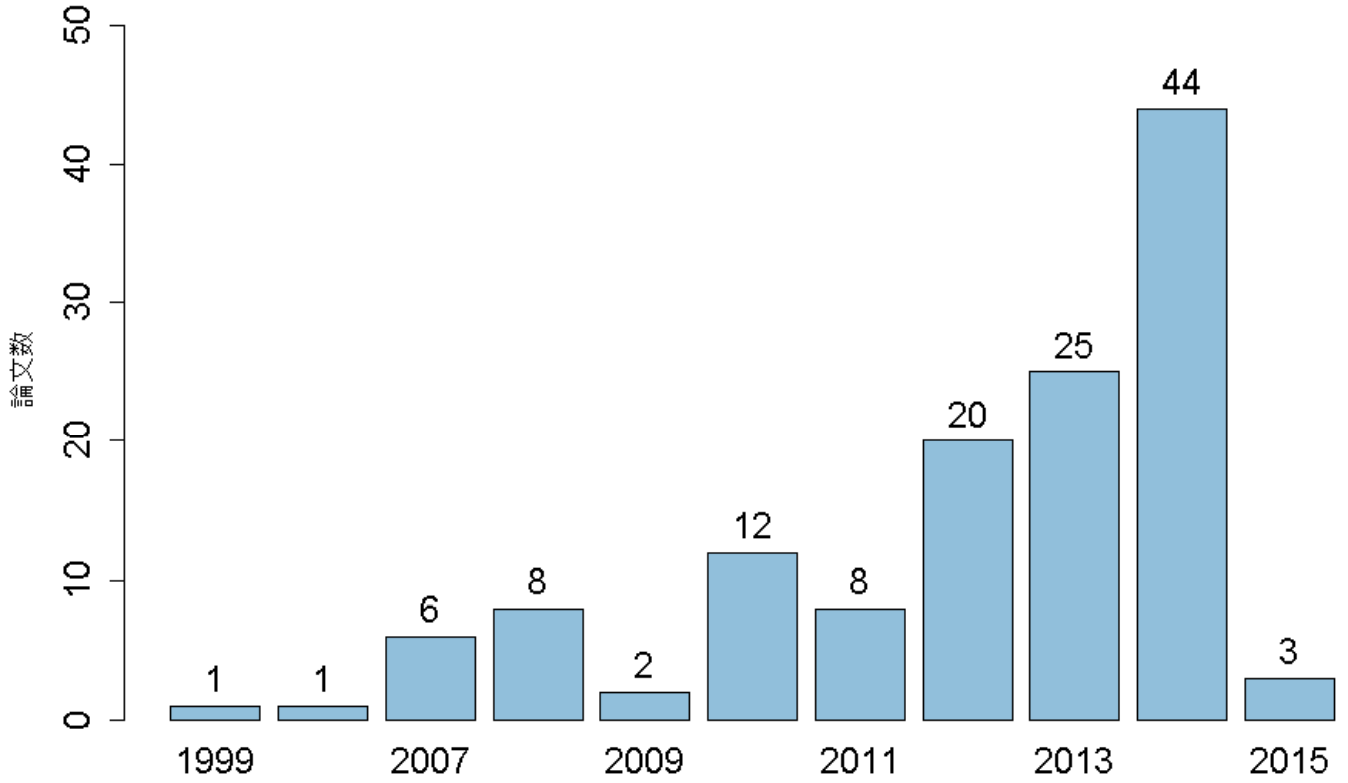
[1] Gayat E et al: Intensive Care Med. 2010 Dec;36(12):1993-2003.
[2] Stürmer T et al: J Clin Epidemiol. 2006 May;59(5):437-47.
[3] Weitzen S et al: Pharmacoepidemiol Drug Saf. 2004 Dec;13(12):841-53.

共変量の数

研究	第1四分位	中央値	第3四分位
[1] 集中治療 2006-2009 (47編)	9	15	22
[2] 全体 1983-2003 (177編)	10	17	28
[3] 全体 2001 (47編)	8	17	27

[1] Gayat E et al: Intensive Care Med. 2010 Dec;36(12):1993-2003.
[2] Stürmer T et al: J Clin Epidemiol. 2006 May;59(5):437-47.
[3] Weitzen S et al: Pharmacoepidemiol Drug Saf. 2004 Dec;13(12):841-53.

日本発の論文数増加 (130編)



検索日: 2015/01/12
検索式: (propensity score*[tiab] OR propensity match*[tiab] OR propensity analy*[tiab]) AND (japan*[tiab])

57

雑誌別の日本発論文数 (130編)

順位	雑誌名	論文数
1	Circ J	8
2	Int J Cardiol	4
3	Cardiovasc Diabetol	3
4	Eur J Cardiothorac Surg	3
5	Cardiovasc Interv Ther	2
6	Clin Ther	2
7	Crit Care	2
8	Crit Care Med	2
9	Eur Heart J	2
10	Hepatol Res	2
11	Int J Hematol	2
12	J Atheroscler Thromb	2
13	J Cardiol	2
14	J Gastroenterol	2
15	J Gastroenterol Hepatol	2
16	J Thromb Haemost	2
17	JAMA	2
18	PLoS One	2
19	Value Health	2
20	その他 (1編)	82

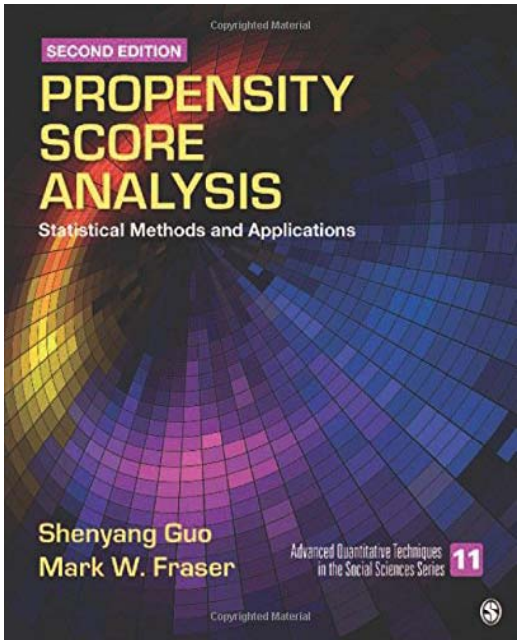
傾向スコアの将来展望

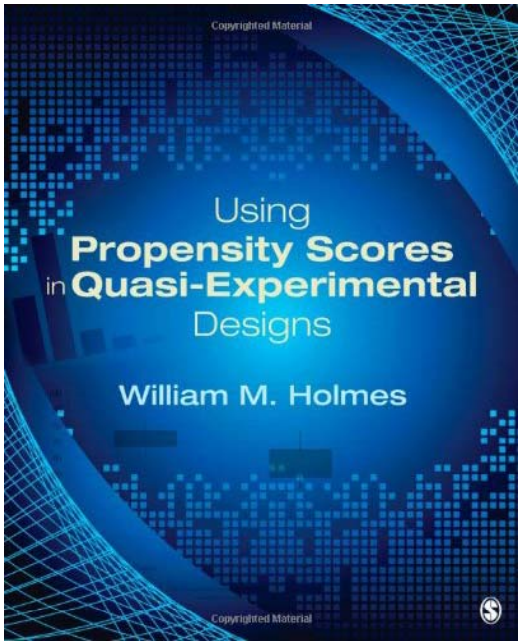
■データベースの整備
- ◆未測定の共変量を減らす
- ◆欠測発生を予防する

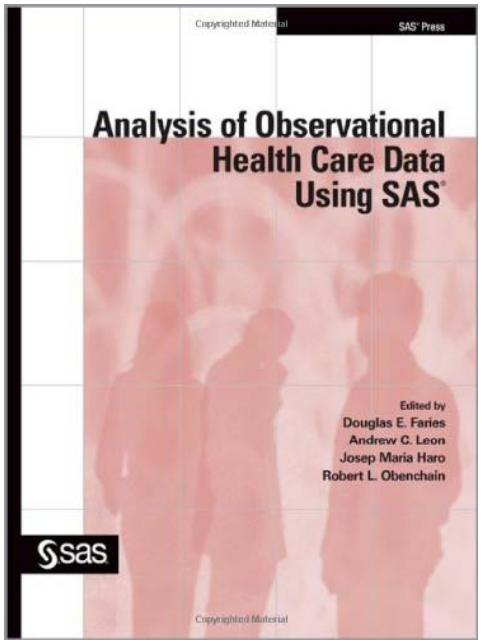
■統一ガイドラインの整備
- ◆実施と報告のガイドラインが必要
- ◆複数の利害関係者により作るべき

Borah BJ et al: J Comp Eff Res. 2014 Jan;3(1):63-78.

入門書







60

導入論文

Statistical Science
2010, Vol. 25, No. 1, 1–21
DOI: 10.1214/09-STS313
© Institute of Mathematical Statistics, 2010

Matching Methods for Causal Inference: A Review and a Look Forward

Elizabeth A. Stuart

Abstract. When estimating causal effects using observational data, it is desirable to replicate a randomized experiment as closely as possible by obtaining treated and control groups with similar covariate distributions. This goal can often be achieved by choosing well-matched samples of the original treated and control groups, thereby reducing bias due to the covariates. Since the 1970s, work on matching methods has examined how to best choose treated and control subjects for comparison. Matching methods are gaining popularity in fields such as economics, epidemiology, medicine and political science. However, until now the literature and related advice has been scattered across disciplines. Researchers who are interested in using matching methods—or developing methods related to matching—do not have a single place to turn to learn about past and current research. This paper provides a structure for thinking about matching methods and guidance on their use, coalescing the existing research (both old and new) and providing a summary of where the literature on matching methods is now and where it should be headed.

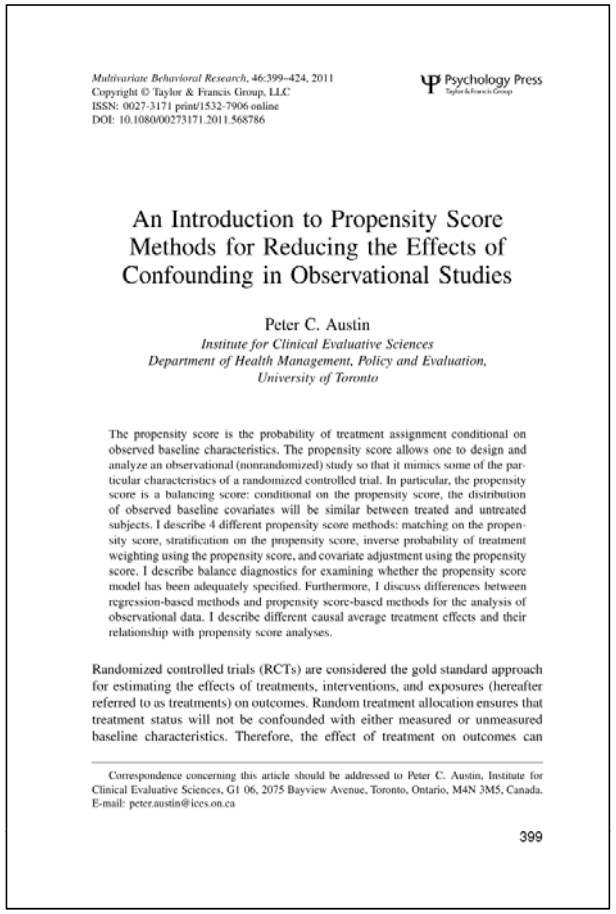
Key words and phrases: Observational study, propensity scores, subclassification, weighting.

1. INTRODUCTION

One of the key benefits of randomized experiments for estimating causal effects is that the treated and control groups are guaranteed to be only randomly different from one another on all background covariates, both observed and unobserved. Work on matching methods has examined how to replicate this as much as possible for observed covariates with observational (nonrandomized) data. Since early work in matching, which began in the 1940s, the methods have increased in both complexity and use. However, while the field is expanding, there has been no single source of information for researchers interested in an overview of the methods and techniques available, nor a summary of advice for applied researchers interested in implementing these methods. In contrast, the research and resources have been scattered across disciplines such as statistics (Rosenbaum, 2002; Rubin, 2006), epidemiology (Brookhart et al., 2006), sociology (Morgan and Harding, 2006), economics (Imbens, 2004) and political science (Ho et al., 2007). This paper coalesces the diverse literature on matching methods, bringing together the original work on matching methods—of which many current researchers are not aware—and tying together ideas across disciplines. In addition to providing guidance on the use of matching methods, the paper provides a view of where research on matching methods should be headed.

We define "matching" broadly to be any method that aims to equate (or "balance") the distribution of covariates in the treated and control groups. This may involve 1 : 1 matching, weighting or subclassification. The use of matching methods is in the broader context of the careful design of nonexperimental studies (Rosenbaum, 1999, 2002; Rubin, 2007). While extensive time and effort is put into the careful design of randomized experiments, relatively little effort is put into the corresponding "design" of nonexperimental studies. In fact, precisely because nonexperimental studies do not have the benefit of randomization, they require

Elizabeth A. Stuart is Assistant Professor, Departments of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA (e-mail: estuart@jhsph.edu).



Multivariate Behavioral Research, 46:399–424, 2011

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2011.568786

Psychology Press
Taylor & Francis Group

An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies

Peter C. Austin
Institute for Clinical Evaluative Sciences
Department of Health Management, Policy and Evaluation, University of Toronto

The propensity score is the probability of treatment assignment conditional on observed baseline characteristics. The propensity score allows one to design and analyze an observational (nonrandomized) study so that it mimics some of the particular characteristics of a randomized controlled trial. In particular, the propensity score is a balancing score: conditional on the propensity score, the distribution of observed baseline covariates will be similar between treated and untreated subjects. I describe 4 different propensity score methods: matching on the propensity score, stratification on the propensity score, inverse probability of treatment weighting using the propensity score, and covariate adjustment using the propensity score. I describe balance diagnostics for examining whether the propensity score model has been adequately specified. Furthermore, I discuss differences between regression-based methods and propensity score-based methods for the analysis of observational data. I describe different causal average treatment effects and their relationship with propensity score analyses.

Randomized controlled trials (RCTs) are considered the gold standard approach for estimating the effects of treatments, interventions, and exposures (hereafter referred to as treatments) on outcomes. Random treatment allocation ensures that treatment status will not be confounded with either measured or unmeasured baseline characteristics. Therefore, the effect of treatment on outcomes can

Correspondence concerning this article should be addressed to Peter C. Austin, Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5, Canada. E-mail: peter.austin@ices.on.ca

[1] Stuart EA: Stat Sci 25:1-21, 2010 .
[2] Austin PC: Multivariate Behav Res 46: 399-424, 2011.

62

63

Nearest neighbor matching

■距離が近いペアを作る

ID	PS
t1	0.48
t2	0.97
t3	0.69
t4	0.68
t5	0.96
t6	0.34
c1	0.31
c2	0.00
c3	0.74
c4	0.02
c5	0.52
c6	0.29

曝露群		非曝露群		
ID	PS1	ID	PS2	|PS1-PS2|
t1	0.48	c5	0.52	0.04
t2	0.97	c3	0.74	0.23
t3	0.69	c1	0.31	0.37
t4	0.68	c6	0.29	0.39
t5	0.96	c4	0.02	0.93
t6	0.34	c2	0.00	0.34
				2.31

距離のレンジは
0.04~0.93

距離の合計は
2.31

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

Optimal matching

■距離の合計が最小になるペアを作る

ID	PS
t1	0.48
t2	0.97
t3	0.69
t4	0.68
t5	0.96
t6	0.34
c1	0.31
c2	0.00
c3	0.74
c4	0.02
c5	0.52
c6	0.29

曝露群		非曝露群		
ID	PS1	ID	PS2	|PS1-PS2|
t1	0.48	c6	0.29	0.19
t2	0.97	c4	0.02	0.95
t3	0.69	c5	0.52	0.16
t4	0.68	c2	0.00	0.68
t5	0.96	c3	0.74	0.21
t6	0.34	c1	0.31	0.03
				2.22

距離のレンジは
0.03~0.95

距離の合計は
2.22

Austin PC: Multivariate Behav Res 46: 399-424, 2011.

65

Take Home Messages

- ルーチンで使う尺度を決めよう
- COMETでコアアウトカムを検索しよう
- COSMINによる尺度の評価研究を探そう
- 変化量の解釈はMDCとMIC
- MDCは安定した状態の2時点で測定
- MICは評価時にアンカーを併せて測定
- リアルワールド臨床データベースを作ろう