〈総説〉

傾向スコアを用いた共変量調整による因果効果の推定と臨床医学・疫学・薬学・公衆衛生分野での応用について

星野崇宏[1)]，岡田謙介[2)]

東京大学　教養学部・大学院総合文化研究科[1)]　東京大学　大学院総合文化研究科／日本学術振興会[2)]

Estimation of Causal Effect Using Propensity Score Methods in Clinical Medicine, Epidemiology, Pharmacoepidemiology and Public Health; A Review

Takahiro HOSHINO[1)], Kensuke OKADA[2)]

[1,2)] Department of Life Sciences, The University of Tokyo
[2)] Japan Society for Promotion of Science

抄録

要旨：医学研究の中でも基礎医学と異なり保健医療や疫学では，無作為割付を伴う実験研究を行うことが難しい．従って，従属変数と独立変数（条件への割付）どちらにも影響を与える共変量や交絡要因の情報を用いた共変量調整を行う必要があり，特に傾向スコアを用いた共変量調整法は近年応用研究で非常によく利用されてきている．本総説では，傾向スコアを用いた共変量調整法の説明を行い，既存の共変量調整法との比較，欧米一流誌での応用例のレビュー，共変量の選択についての議論について概説を行った．

キーワード：観察研究，交絡による偏り，因果推論，セミパラメトリックモデル，共変量選択法

Abstract

In epidemiology and public health, it is not easy to conduct randomized experimental studies. Therefore the information of covariates and confounding factors affecting both independent and dependent variables should be utilized to adjust bias. Propensity score adjustment has been one of the most widely employed covariate adjustment methods in applied researches. In this article, we explained the method of covariate adjustment using the propensity score, compared it with existing methods, reviewed applied researches published in top journals, and discussed the covariate selection problem.

Keywords : observational study, confounding bias, causal inference, semi-parametric modeling, variable selection of covariates

1．はじめに

医学の中でも基礎研究，特に動物を被験体とした実験研究では，例えばラットを無作為に実験群と対照群に割付け，新しい治療方法の有用性を検証するといった無作為割付(random assignment)を伴う実験研究が行われる．

しかし，ヒトを対象にする疫学研究では，倫理的な問題や実行可能性の点から，関心のある曝露要因や治療方法についての被験者の無作為割り付けを伴う実験研究（又は介入研

〒110-0001　東京都目黒区駒場3-8-1
東京大学大学院総合文化研究科広域科学専攻生命環境科学系認知行動科学講座
3-8-1 Komaba, Meguro-ku, Tokyo 110-0001, Japan.

究）を行うことが一般的に不可能である．

また，臨床研究においても無作為化比較試験（Randomized Controlled Trial，RCT，丹後，[1] などを参照）を行うことは患者の抵抗を伴うことが多く，ホーソン効果など，外的な（研究者による）割付の心理学的な効果の問題なども指摘されている．したがって疫学や臨床研究では，研究者による独立変数（要因・条件）の操作を伴わない，いわゆる「観察研究（observational study）」が行われることが多い（本総説では，無作為割付による純粋な実験研究以外を一括して観察研究と呼ぶ）．

一般に観察研究によって独立変数の従属変数（結果変数）に対する影響を調べる際には，従属変数に影響を与える共変量（剰余変数，または交絡変数・交絡要因とも言うが，以後共変量という用語を用いる）の分布が独立変数の値によって異なる（＝交絡する）可能性がある．

例えばヒトを対象にした喫煙（＝独立変数）の大腸がん発症（＝従属変数）リスクを考える際には，喫煙の有無を研究者が操作することは出来ないために，あくまで「喫煙群」と「非喫煙群」での「XX 年間での大腸がんの発症率」を調べることになる．しかしこのような研究デザインでは，喫煙にも発ガンにも関係する飲酒量などの共変量の影響が除去されないために，喫煙の発ガンへの単独の効果（＝因果効果）を知ることが出来ない．

そこで共変量の影響を除去するために，これまでも共分散分析などの様々な統計解析が利用されてきたが，共分散分析的な手法は従属変数と共変量の関係を事前に線形関数などと指定する必要があるなど様々な制約が多いという欠点を有することから，Rosenbaum & Rubin[2] が提案した概念である傾向スコア（Propensity score）を利用した共変量調整法が近年応用研究に利用されるようになり，注目を集めている．

具体的な解析例としては，例えば臨床医学や疫学ではアスピリンの冠動脈疾患に対する有用性（Gum ら，[3]）やフェノバービタルの曝露による知能への影響（Reinisch, Sanders, Mortensen & Rubin, [4]），退院後の治療のタイプによる心筋梗塞の予後への影響（Ayanian ら，[5]），非定型的抗精神病薬（atypical antipsychotic medications）の脳血管疾患副作用，認知機能低下，死亡率に関するリスク（Wang ら，[6]）などの解析がなされている．

また，医療政策に関しては復員軍人庁健康局（Veterans Health Administration）の病院の治療の質に関して（Petersen, Normand, Daley & McNeil, [7]），退院後のケアプログラムについて（Coyte, Young & Croxford, [8]）などの応用例が見られる．

さらに評価研究といわれる領域でも，養育環境に問題のある幼児に対するケアの効果（Hill, Waldfogel & Gunn, [9]），アメリカ麻薬管理局（The United States Office of National Drug Control Policy）による全米での対麻薬キャンペーンの効果（Lu, Zanutto, Hornik, & Rosenbaum, [10]），ボリビアの社会投資ファンドによる公衆衛生分野での社会資本へのインパクト評価（Newman, Pradham, Rawlingsm, Ridder, Coa & Evia, [11]），銃暴力への曝露の将来の暴力的行動への因果効果の推定（Bingenheimer, Brennan & Earls, [12]）などが行われている．

しかし，これまで傾向スコアによる調整法は日本において「紹介されなかった多変量解析法」（佐藤，[13]）であり，欧米の一流誌で非常によく利用されているわりには国内の研究者による利用がほとんどなされていなかった．

そこで本稿では傾向スコアを用いた「因果効果」（causal effect：Rubin, [14]）または「平均処遇効果」（treatment effect：Neyman, [15]）の推定に関する概説と医学研究，特に公衆衛生分野・疫学・薬学における応用例の紹介を行う．

本稿では傾向スコアを用いた解析法の数理的な部分の説明について詳細には記載しないので，詳しくは星野・繁桝，[16]，狩野，[17] を参照いただきたい．

本稿の構成は以下の通りである．第 2 節では因果効果の定義と無作為割付の重要性について説明する．第 3 節は既存の調整法とその問題点について述べる．

第 4 節では傾向スコアの定義と解析のための条件について述べる．第 5 節では現在複数提案されている傾向スコアを用いた具体的な解析法のうち，代表的な手法についての説明を行う．第 6 節では傾向スコアによる調整が一定の条件の下では共分散分析的な手法より優れている点を明示する．

第 7 節では傾向スコアが利用可能であるための前提条件である「強く無視できる割り当て」条件のチェック法と共変量の選択について議論する．

そして第 8 節では傾向スコア解析の具体的な利用例を，特に欧米の一流誌に掲載された解析例を中心に提示し，代表的な応用研究について様々な観点からレビューを行う．第 9 節では傾向スコア解析法のいくつかの拡張について概説する．最後に簡単なまとめを提示する．

2．因果効果の定義と無作為割付の重要性

無作為割付を伴った実験研究は一般に，「研究者にとって関心のある要因の効果のみを知ることが出来る」つまり内的妥当性（Internal validity，Cook & Campbell, [18]）があるとされる．これに対して，研究者による独立変数（割付）の操作性がない観察研究や，実験研究であっても無作為割付を伴わない研究は，内的妥当性が低く，関心のある要因の効果のみを知ることは難しいとされる．

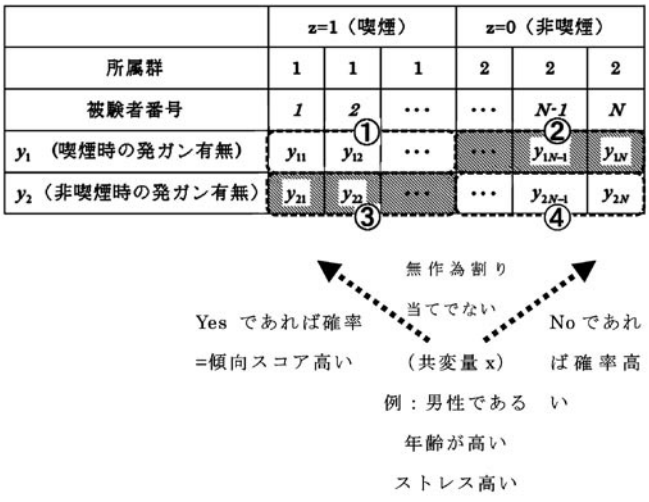
このことは，割付を欠測データの問題（Little & Rubin, [19]，Rubin, [20]）として考えると明確になる．

ここである従属変数（例えば 1 年後の患者の生存率）が 2 つの条件（例えば放射線化学療法と手術）間でどのように異なるかを調べることに関心があるとする．

ある被験者が条件 1 に割付られた時の従属変数の値を $\boldsymbol{y}_1$，同様に条件 2 の時の従属変数の値を $\boldsymbol{y}_2$ とする．また被験者が条件 1 なら $\boldsymbol{z}=1$，条件 2 なら $\boldsymbol{z}=2$ の値をとる変数 $\boldsymbol{z}$（これを割付変数と呼ぶ）を導入する．

また，条件 1 に割付られた被験者の集団を群 1，同様に条件 2 に割付られた被験者の集団を群 2 とする．

ここで注意すべきは，群１の被験者については y_1 しか観測されないが，本来は y_2 も存在し，これを観測することはできない（欠測である）と考えているということである（群２についても同様）（図１参照）．



＊網掛け部分は実際は得られないデータの部分

無作為割り当てなら無視してよいが、観察研究では無視できない

図１　：無作為でない割り当てをデータの欠測の問題として考える

このとき群１と群２の「因果効果」または「平均処遇効果」を d_{12} を

d_{12} ＝「y_1 の周辺期待値」－「y_2 の周辺期待値」

と定義する．ここで「y_1 の周辺期待値」，「y_2 の周辺期待値」とは，治療法の例で言えば，それぞれ「全ての患者が放射線化学療法を受けた場合の生存率」，「全ての患者が手術を受けた場合の生存率」を意味する．したがって d_{12} は患者の状態や病院の設備などの共変量・交絡要因の分布が２群で異なることによる影響が除去された「治療方法単独の生存率への効果」であるといえる．

ここで群１と群２の周辺期待値はそれぞれ

$\frac{1}{N}\sum_{\text{全被験者}} y_1$（＝図1の”①+②”の平均）

$\frac{1}{N}\sum_{\text{全被験者}} y_2$（＝”③+④”の平均）

によって正しく推定（不偏推定）できる（但し N は群１と群２を合わせたサンプルサイズ），実際には，群１の被験者の y_2 や群２の被験者の y_1 は観測されないので，上記の計算をすることは不可能である．

しかし，無作為割付が行われている実験研究では d_{12} の推定を，観測されているデータから

$\frac{1}{N_1}\sum_{\text{群1の被験者}} y_1 - \frac{1}{N_2}\sum_{\text{群2の被験者}} y_2$（＝図１の①の平均と④の平均の差）

によって行うことができる（ただし N_1, N_2 はそれぞれの群のサンプルサイズ）．なぜなら，無作為割付ならばどちらの群に属するかは「無作為」であり２つの群は条件の違いを除けば等質なので，人数が多くなるにつれて $\frac{1}{N}\sum_{\text{全被験者}} y_1$ ，$\frac{1}{N_1}\sum_{\text{群1の被験者}} y_1$ 、$\frac{1}{N}\sum_{\text{全被験者}} y_2$ と $\frac{1}{N_2}\sum_{\text{群2の被験者}} y_2$ が一致するからである．

しかしこの関係は無作為割付を伴わない観察研究では成立しないため，この方法で d_{12} の推定を行うことはできない．

逆に言えば，この因果効果を観察研究でも推定できれば，無作為割付による実験で得られた結果と同様に因果効果の推定ができると考えられる．逆にヒトを対象とする研究領域においては，研究者の関心のある要因が操作可能ではなく，理論的に無作為割付不可能な場合があること，倫理的な理由で無作為化が適切でない場合などがある．さらに，心理学で明らかにされているホーソン効果など，無作為割付による実験研究という状況は非常に不自然になることが多く，結果の生態学的妥当性（Ecological validity）を欠く恐れがある（Cook & Campbell,[18] などを参照）ことから，観察研究での因果効果を推定することが様々な分野で求められている．

さて，Rubin の因果効果の定義を利用する意義を，疫学でよく引用される Hill,[21] の因果関係の判断基準に照らして考えると，Hill では

１：強い相関関係がある．
２：相関関係が常に成立する．
３：相関関係に特異性がある．
４：時間的前後関係が明確．
５：現象の背後にメカニズムが想定できる．
６：もっともらしい．
７：首尾一貫している（これまでの他の知見と矛盾しない）．
８：実験的な証拠がある（独立変数に操作性がある）．
９：アナロジーが成立する．

といった条件が成立する場合に因果関係が存在するとしているが，ここでは特に8についての基準を緩和するために，独立変数の操作性を仮定せずに交絡要因の影響の除去を目指すということになる．

3．既存の共変量調整法とその問題点

観察研究の多くでは従属変数と独立変数のみを測定するのではなく，それらに影響を与えるであろう共変量も測定し，これを考慮した解析が行われる．このような方法は複数存在するが，大きく分けて３つに分類することができよう．つまり，(1) 均衡化：共変量の値が同じになるペアを作ることで２つの群の被験者を構成する，(2) 恒常化・限定：共変量のある値の被験者のみに限定して解析を行う，(3) 統計的調整（マッチング・層別解析・共分散分析），である．(1) と (2) は研究デザインによる調整，(3) は統計解析に

よる調整と考えることができるが，これらの方法もそれぞれ欠点を有している．例えば「均衡化」では，共変量に連続変数が存在するときは，一般に両群で完全に値が一致するようなペアを作ることはできない．そこで，なるべく近い被験者をペアにする必要があるが，その方法が恣意的である．また，共変量の数が多いと実際に行うことは無理になる．また「恒常化・限定」では得られた知見が研究を行った共変量の値においてのみに限定されてしまい，研究の一般化可能性が低減する．「マッチング・層別解析」ではどのようにマッチング・層別するかに関して恣意性が残る．また，共変量の数が多いと実際に行うことは無理になる．

したがって，応用研究で最もよく利用される手法は，共分散分析的な手法である．これは，共変量も同時にモデルに含めて解析する手法であり，その最も単純なものとして共分散分析があるが，ここではCox回帰やパス解析，グラフィカルモデリングや構造方程式モデリングなども含めてより広義に考える．

共分散分析的手法の最大の問題点は，従属変数と共変量の関係をモデル化する必要があるということである．このモデル化では回帰関数を間違って指定すると（例えば二次関数の関係があるのに線形と仮定すると）誤った結果が導かれることなどがよく知られている．また，共分散分析的手法で推定することができる回帰係数自体は因果効果 と等しくはならない．

このように既存の共変量調整法は様々な問題点を有しており，これらの問題点を解決する新しいタイプの（統計学的な用語でいうところの“セミパラメトリックな”）共変量調整法として，傾向スコアを用いた解析法が近年医学や経済学など様々な分野において利用されてきている．

4．解析の前提条件と傾向スコアの定義

無作為割付が不可能な観察研究において，因果効果を推定する方法として，Rosenbaum & Rubin,[2]は，傾向スコアという新しい概念を提案した．これは，複数の共変量を一つの変数に集約することができれば，その一変数の上でマッチングや層別化などを行うことができ，前節のような問題が起こらない，ということから考え出された概念である．

傾向スコアを定義する前に，まずは傾向スコアを用いて因果効果を推定できるための前提条件であるところの「強く無視できる割り当て」(Strongly Ignorable Treatment Assignment) 条件を説明する．

「強く無視できる割り当て」条件

共変量を所与とするときに「強く無視できる割り当て」であるとは，「どちらの群に割付られるかは観測された共変量の値に依存し，従属変数の値の高低によっては依存しない」（より数理的な表現としては星野・繁桝,[16]を参照）という条件のことである．

この条件は非常に強く感じられるが，既存の調整法であるマッチングや層別解析を行う場合も，実はこの仮定を行っている（Rosenbaum & Rubin,[22]）ことに注意するべきである．また，割付は従属変数の測定より時間的に先行しているので，従属変数の値によって割付が決まるということはあまりありえない．従ってこの条件の重要な点は「観測された共変量」によって割付を説明できなくてはならない，観測されていない共変量・交絡要因が割付に影響を与えていない，という点である（この問題については第7節参照）．

傾向スコアの定義（Rosenbaum & Rubin,[2]）

第i被験者の共変量ベクトルを$\boldsymbol{x}_i$，割付変数の値をz_iとするとき，群1へ割付られる確率$e_i = Pr(z_i = 1 | \boldsymbol{x}_i)$を第i被験者の傾向スコアという．

実際には各被験者の傾向スコアの真値はわからないので，データから推定する必要があり，推定においては共変量を用いて割付を説明するモデルとしてロジスティック回帰分析モデルが使用されることが多い．傾向スコアのより理論的な説明については，狩野,[17]，佐藤・松山,[23]を参照されたい．

5．傾向スコアを用いた具体的な解析方法

傾向スコアを用いた調整は，前提条件である「強く無視できる割り当て」条件が満たされていれば，共変量全てを用いて調整を行ったのと同じだけ偏りを減少させることができ，（2節で説明したように群1の被験者の や群2の被験者の$\boldsymbol{y}_1$は観測されないにも関わらず）傾向スコア及び観測されている従属変数（つまり群1の被験者の$\boldsymbol{y}_1$と群2の被験者の$\boldsymbol{y}_2$）の情報を用いれば因果効果を推定することができる（詳しい理論的説明は星野・繁桝,[16]を参照）．

傾向スコアを用いた調整法はすべて二段階推定法であり，以下の2つのステップを踏む必要がある．

1）傾向スコアの推定

割付変数を共変量によって説明するモデルを設定し，そのモデルの母数の推定を行う．母数の推定値を用いて，各被験者に対して「条件1に割付られる予測確率」を計算し，これを傾向スコアの推定値とする．

一般にロジスティック回帰モデルが利用されることが多いが，ノンパラメトリック回帰によって母数推定をせずに直接予測確率を計算している例も多い．

2）推定された傾向スコアを用いた調整

上記で推定された傾向スコアを用いて，具体的な調整を行う方法としてRosenbaum & Rubin,[2]はマッチングと層別，共分散分析の3つの方法を提案しており，これまでの解析例の多くではこれらの方法が利用されてきた（8節の解析例を参照）．しかし現在では後述する重み付け平均を用いた方法に関する理論的研究が進み，しだいに利用例が増えつつある．

傾向スコアを利用して因果効果d_{12}の推定を行う方法はこれまでにいくつか提案されているが，Rosenbaum & Rubin,[2]が提案したのは以下の3つの方法である．

(1) マッチング

2つの群で傾向スコアが等しいと見なせる被験者をペアにして，その差の平均を推定値とする．

ここで，マッチングを行う際に傾向スコアの差がなるべく小さくなるようにマッチングを行う方法は複数ある．また，差が小さいペアを構成できなくなった時点でマッチングをやめるということがしばしば行われる．マッチングによる傾向スコア解析についてはRosenbaum,[24]も参照されたい．

(2) 層別解析

傾向スコアの大小によっていくつかのサブクラスに分け，その各クラスで2つの群の平均を算出し，それらを併合した全体としての効果の推定量を計算する（さらに詳しい議論はRosenbaum & Rubin,[25]参照）．

(3) 共分散分析

傾向スコアを共変量とした共分散分析を行う．

複数の共変量を用いてそのままマッチング・層別することは事実上不可能であるが，傾向スコアを用いれば共変量を1次元に縮約し，その上でマッチングや層別解析を行うことができるので，非常に有用である．

既存のマッチング法との比較に関して，Rosenbaum & Rubin,[26]では既存のマッチングによる推定の偏りを以下の3つの原因の和に分けることができるとした．

つまり，(i)「強く無視できる割り当て」条件からの逸脱（これについては後述）(ii) 実験群の被験者に対応する対照群の被験者が見つからない問題 (iii) 不正確なマッチングによる影響，である．

同研究では傾向スコアを用いたマッチングでは (ii) の問題が解決され，かつ (iii) の影響を減少させることが指摘されている．

しかし，上記にあげた3つの手法には以下のような欠点がある．

(i) いずれの方法でも，3群以上の比較に関心がある場合は2群ごとに別々の傾向スコアを推定する必要があるために，因果効果を求めるための母集団が各2群の解析ごとに異なってしまう（詳しくは第9節参照）．

(ii) マッチング・層別解析では因果効果の推定値は計算できるが，その標準誤差が正確に計算できない．従って統計学的に正しい検定も行えない（但し，傾向スコアを推定したことによる影響を無視した，単純なマッチング・層別解析による因果効果の検定は応用研究ではよく行われている）．

(iii) マッチング・層別解析ともに，従属変数の周辺期待値の推定ができない．

(iv) マッチングの方法は一意に決まらないので恣意性が残る．層別においても，Rosenbaum & Rubin,[24]は5層以上とればよいと提案し，実際に多くの研究で5層に層別がされているが，これが適切であるという証拠は明確ではなく，層別の基準が恣意的である．

(v) 通常行われている1：1マッチングでは，被験者の数が多い群でデータの多くが無駄になる．特に問題なのは，被験者数が少ない方の群の共変量の分布の上で期待値を取ったときの因果効果の推定になってしまう．

(vi) 共分散分析のモデルで傾向スコア解析を行うための前提条件として，傾向スコアと目的変数が線形な関係にある必要があるが，そのような関係を仮定が成立するかは分からない．

Rosenbaum & Rubin,[2]で提案された3つの手法には上記のような欠点があり，その後以下の3つの方法が提案され，次第に利用されつつある．

(4) Horovitz–Thompson 型推定量

Rubin,[27]，Rosenbaum,[28]らは層別標本抽出におけるHorovitz & Thompson,[29]の方法を拡張した「傾向スコアによる重み付け推定法」を提案している．

これは，傾向スコアの関数による重み付け平均によって，「$\boldsymbol{y}_1$の周辺期待値」と「$\boldsymbol{y}_2$の周辺期待値」を推定する方法であり，具体的にはそれぞれ，

$$\frac{\sum_{i:\text{全被験者}} \frac{z_i}{e_i} y_{1i}}{\sum_{i:\text{全被験者}} \frac{z_i}{e_i}} \quad \frac{\sum_{i:\text{全被験者}} \frac{1-z_i}{1-e_i} y_{2i}}{\sum_{i:\text{全被験者}} \frac{1-z_i}{1-e_i}}$$

によって推定される．但し$\boldsymbol{y}_{1i}$は被験者iが条件1に割付られた場合の従属変数の値であり，この被験者が群1に所属（$\boldsymbol{z}_i=1$）していれば値が観測され，$\boldsymbol{y}_{2i}$は被験者iが条件2に割付られた場合の従属変数の値であり，この被験者が群2に所属（$\boldsymbol{z}_i=0$）していれば値が観測される．また，これらの差が因果効果の推定値となる．

(5) Rotnitzky らの重み付き一般化推定方程式

上記の傾向スコア解析の目的は，無作為割付がなされたときの周辺期待値や因果効果の推定であった．Rotnitzky & Robins,[30]は回帰モデル（より一般的には周辺平均構造）の母数推定のための重み付き一般化推定方程式(Generalized Estimating Equation，Liang & Zeger,[31])法を提案している（詳細は星野・繁桝,[16]を参照）．

(6) 重み付けM推定量

また，筆者らは平均構造の推定だけでなく，一般化線形モデル，構造方程式モデリングや変量効果モデル，階層的モデルにおいても利用可能な下記の方法を提案している（星野,[32]，Hoshino, Kurata & Shigemasu,[33]），

これは（1）傾向スコア算出のモデルの母数推定として最尤法を用いる，(2) 最尤推定値で傾向スコア算出の際のモデルの母数を置き換えることで，各被験者ごとの傾向スコアの推定値を得る，(3) 傾向スコアの推定値の逆数で対数尤度などM推定量(Huber,[34])を与える目的関数の重み付けをし，それを最大化する値を従属変数だけの周辺尤度の母数の推定値とする，という方法であり，正確な標準誤差の推定法や検定手法も提案されている．またこの方法は，傾向スコアの推定値の逆数を各被験者の重みの変数とすればよいので，推定値の計算に関しては既存のソフトで容易に実行できる．

しかし標準誤差の計算や検定については Hoshino，Kurata & Shigemasu で導出された結果を利用する必要がある．

6．傾向スコアによる調整が共分散分析的手法より優れている点

ここで共分散分析とこれらの傾向スコアによる調整法との比較が様々な論文で行われており，傾向スコアの優位な点が指摘されているので，これについて紹介したい．

(1) 傾向スコアは共変量を一変数に縮約しているので，２つの群において共変量の値に重なりがない（または少ない）場合でも利用できる (Rubin,[35])．

(2) 共変量と従属変数のモデル設定を行わなくてもよい
第３節でも論じたが，共分散分析は従属変数と共変量の間に既知の関数関係を想定する必要がある．しかし，傾向スコア的手法ではその必要はない（但し，正しい関数関係が想定できる時には，回帰モデルの設定を行うことで共変量の変動を除去できるため，例えば検定の検出力を向上させることができるなどの利点がある)．
ここで，２節でも説明したように従属変数には欠測があるため，共変量と従属変数のモデルをデータだけから想定することは難しい．また，傾向スコアを用いる場合は傾向スコア推定のための群への割付モデルの仮定が必要であるが，一般に従属変数の次元は多次元であるが割付変数は1次元である．したがって群への割付モデルのほうが共変量と従属変数の回帰モデルよりはモデル指定が容易である，このため，傾向スコアを用いた解析法がよく利用されている．

(3) モデルの誤設定に強い
Drake,[36] はシミュレーション研究を用いて，関心のある要因と共変量（２つ）を共に説明変数としたモデル（従属変数が連続なら共分散分析，２値ならロジスティック回帰）と層別による傾向スコア解析を比較した結果，モデルが正しい場合の推定の偏りはどちらも同じ程度に小さく，共変量を無視したらどちらも同程度偏ること，誤ったモデルで推定した場合は傾向スコア解析の方が偏りが小さいという結果になった．
また，Cepeda，Bostonm Farrar & Storm,[37] では従属変数が２値である場合のシミュレーションを行ったところ，ロジスティック回帰分析に比べて傾向スコアを用いた層別解析はモデルの誤設定に頑健であり，検出力の高い検定が可能であり，さらに生起率が多くない場合ではロジスティック回帰よりバイアスが小さいことがわかった．

このように，傾向スコアを用いた調整法については既存の共分散分析的手法にくらべて様々な利点が指摘されている．

7．前提条件のチェックと共変量の選択方法について

第４節で述べたように，傾向スコアによる調整によって因果効果の推定が可能になるためには「強く無視できる割り当て」条件が成立している必要がある．つまり，この条件が満たされるように共変量を選択する必要があるということである．

しかし，この前提条件が成立することを示すためには，観測できない欠測値を知ることが必要であるので，直接確認することは実際には不可能である．

前提条件が成立していることを間接的にチェックする方法はいくつか提案されているが，実際に応用論文で利用されている２つの方法を以下に示す．

(1) 割付を共変量が説明していることを示す．
傾向スコアを計算するときのモデル（ロジスティック回帰モデル）のフィットが良いことを確かめる．例えば，具体的には擬似決定係数（pseudo-R^2）や c 統計量，モデルによる割付の正判別率が高いかどうかを確かめる（近年の応用例では c 統計量が0，8以上であるということが，医学系の論文誌でのスタンダードになっているようである)．
モデルフィットが良ければ，観測されていない他の共変量の影響が無いと言えるため，間接的に前提条件をチェックできたことになる．

(2) 共変量自体の分布を調整していることを示す．
傾向スコアによる調整を共変量に対して行い，群間で分布の差が消えることを確認する．傾向スコアを推定する時に利用した共変量の群間差を調整することができることは，調整がうまく行えることの前提条件の一つであるからである．
D'Aostino,[38] には傾向スコア算出と共変量自体の調整をチェックするための簡単な SAS プログラムが記載されている．

上記のチェックを用いた結果として，「強く無視できる割り当て」条件が成立していないと考えられる場合には，観測していない共変量・交絡要因 (Unmeasured Confounder) が割付に影響していると考えられる．観測していない交絡要因の影響については例えば Rosenbaum & Rubin,[39] などで考察されている．

このように，共変量の選択と前提条件の成立の成否は直接関連する．

多くの応用研究では，共変量の選択については，共分散分析同様，「理論上または先行研究での知見から，調整を行うべき変数」を投入し，確認のために一番目のチェックが行われ，Rosenbaum など方法論を研究している研究者が応用研究に共著者として参加している場合，二番目のチェックが行われている程度である．

実際，Weitzen ら，[40] は傾向スコアを用いて解析が行われた47の応用研究を共変量の選択がどのように行われているかという観点からレビューしている．

その結果「半数以上の研究で変数選択基準が明記されていない」こと，「大部分の研究で適合度が明記されていない」こと，「c 統計量も半分の研究で書かれていない」ことを示し，傾向スコアを使った研究の多くで「傾向スコア推定のための割付のモデリング」の検証が軽視されていることに警鐘を鳴らしている．

上記に記載したチェックはある程度重要ではあるが，それだけで共変量選択を行うことの問題点も指摘されている．

例えば星野・前田，[41] は「割付（独立変数）をうまく説明するような共変量を選択する」という上述（1）の方法では調整が必ずしもうまく行えないことを指摘し，「従属変数に関連がある共変量の選択」を行う共変量選択法を提案している．

同様に Brookhart，Schneeweiss，Rothmanm，Glynn，Avorn & Stürmer,[42] もシミュレーション研究から，

・割付に強い関連がある共変量よりも，従属変数に強い関連がある共変量を選ぶ方が因果効果の推定の偏りが少なく，かつ推定量の分散が小さくなる（したがって検出力が高くなる）こと

・割付には関連が強くても，従属変数にはあまり関連がない変数を共変量に加えると，推定の偏りはあまり変化しないが，推定量の分散が大きくなってしまい，結果として平均二乗誤差（真値からのズレの指標）が大きくなってしまうことを示している．また具体的には，一般的に傾向スコアを用いた研究で0.8以上が目安とされている c 統計量が，例えば0.67程度であっても，従属変数に関連の強い共変量を選択すれば，十分偏りのない調整が可能である場合があることも示されている．

したがって，上述の（1）の基準を満たすように，割付をよりよく説明する共変量を利用することのみを重視するよりは，「理論上または先行研究での知見から，調整を行うべき変数」を投入し，一応（1）の基準もチェックしてみるのが良いと考えられる．

また，共変量の選択に関してはもう一点重要な議論がある．Rosenbaum,[43] は，無作為割付による実験研究においては処遇の割付の後に測定され，処遇の影響を受ける共変量（Posttreatment variable）による調整はかえって偏りを生むことを示している．共変量の選択に関しては，共変量が従属変数・独立変数より理論的な意味で先行している必要があることに注意する（測定時が実際に先行しているかどうかはともかく，理論的に共変量が従属変数の結果になっていないということに注意する）べきということになる．

8．具体的な解析例について

本節では，医学研究における傾向スコアを用いた共変量調整の解析例を紹介する．

（1）Gum ら，[3] による冠動脈疾患患者に対するアスピリンの有用性の研究

アスピリンが心筋梗塞を減らすことがこれまで示唆されてきたが，これまで直接の研究は無かった．

そこで，冠動脈疾患患者またはその疑いのある6174人に対して1990年から1998年までのコホート研究を行った．このうち，アスピリンを服用しているのは2310人であり，当然ながら服用群と非服用群は無作為に割付られていない．

単純な死亡率の比較からは，アスピリン服用群も非服用群もともに4.5% と差がなかった．

アスピリン服用群と非服用群への群別に関係があると思われる共変量として，年齢や他の治療薬の服用，喫煙や心臓疾患に関する検査指標など34変数を取り上げ，ロジスティック回帰分析を行って傾向スコアを計算した．

傾向スコアを用いたマッチングを行い，マッチングがうまく行かない被験者を分析から外すことで最終的に1351のペアを構成した．

結果として，服用群と非服用群の死亡率はそれぞれ4%，8% となり，大きな差が開くことが分かった．

またこの研究では「強く無視できる割り当て」条件が成立することをロジスティック回帰分析のフィットの指標である c 統計量が0.83と高いことで確認している．

（2）Coyte ら，[8] による退院後のケアのタイプの評価

この研究の目的は関節置換（joint replacement）手術後，患者が退院してどこに収容されるかによって，再入院率や病気のケアのための総費用がどれくらい異なるかを知ることにある．

この場合，当然ながら退院後の患者を無作為に退院後のケアの各タイプに割付ることはできない．そこで，退院後にどこに収容されるかを予測する共変量として患者の年齢，性別，合併症，住居が都市かどうかなどの９つの変数を用いて傾向スコアを計算し，傾向スコア上で５つの層に分けて解析を行った．

傾向スコア算出の際はいくつかの変数の交互作用を含めてモデリングを行っている．ここでは独立変数である「退院後のケア」のタイプは４つあり，リハビリテーション病院に入院した後に家にもどり，ホームケアサービスを受けない群（RS），リハビリテーション病院に入院した後に家にもどり，ホームケアを受ける群（RH），そのまま家に戻りホームケアを受ける群（HC），家に戻りホームケアを受けない群（SC）である．

傾向スコアによる調整の結果，RS 群の方が RH 群よりも却って再入院率が低く，かつコストが低いことが示された．

（3）Wang ら，[6] による非定型抗精神病薬のリスクに関する研究

FDA が「非定型抗精神病薬は高齢者の死亡率を高める」という発表をしたが，既存の抗精神病薬投与群との比較を行っていなかった．

そこで後ろ向きコホート研究（症例対照研究）のデータを用いて，非定型抗精神病薬および既存の抗精神病薬間の死亡率の差を調べた．年齢・性別・様々な病気の有無・他の薬の利用などの25変数を説明変数としたロジスティック回帰分析を用いて傾向スコアを算出した．また，「強く無視できる割り当て」条件のチェックには c 統計量（ここでは0.845）を用いている．

結果として調整前だけでなく，傾向スコアを用いた調整後でも非定型抗精神病薬投与群の方が死亡率は低かった．また痴呆の有無，介護ホームへの入居の有無で分類した後でも同様の結果となった．したがって非定型抗精神病薬が既存の薬よりも望ましくないとのFDAの見解は正しくなく，むしろ前者を利用するべきである，と結論している．

(4) McWilliams ら，[44] による保険加入の健康診断の受診率への影響

米国では日本のような国民皆保険制度がなく，保険の加入は個人に委ねられている．保険に加入していない成人は相対的に適切なケアを受ける機会に恵まれず，健康上の不利な影響を受けることが知られていたが，「保険加入の有無」も「医療機関への受診」も個人の裕福度や健康状態に影響を受けるため，単純に加入群と非加入群の比較をすることには意味が無い．

そこでMcWilliamsらは保険加入がもたらす各種健康診断の受信率への影響を調べた．ミシガン大学社会調査研究所が公開している「健康と退職に関するパネル調査」データの分析を進めた結果，保険加入群と非加入群の間では，傾向スコアによって社会人口学的変数の影響を調整してもコレステロール検査，マモグラフィ（女性），前立腺検査（男性）の受信率にはっきりと差があることがわかった．

(5) Shishehbor ら，[45] による社会経済的地位の死亡率に及ぼす影響の研究

社会経済的地位（Socio-Economic Status，SES）が低いことは心臓血管系リスクや死亡率を高めることが知られているが，その媒介経路は明らかではない，Shishehborらはどの生理学的特性がSESと死亡率との連関を説明するのかを，1990年から2004年にわたる縦断研究により調べた．患者はSES得点を使って4群に分けられたが，傾向スコア・マッチングにより年齢や性別などの共変量はよく調整できた．この結果，SES得点が低いことは生活機能の阻害や心拍数回復の異常と独立に関連していることがわかり，またSESの低さは死亡率を有意に予測した．低SES群の患者でもこうした臨床的特徴を改善する努力をすることにより，死亡率を下げることができる可能性がある．

(6) Stenestrand ら，[46] の血行再建術についての研究

急性冠症候群における血行再建術の転帰についての無作為化試験には，相対する結果の先行研究が知られていた．Stenestrandらはスウェーデン国家死亡記録（Swedish National Cause of Death Register）の2次データのうち心臓部門のある61の病院のデータを用い，血行再建術を行った群（n=2.554）と行わない群（n=19.358）とを比較して1年以内の死亡率を調べた（調査期間1995年～1998年）．血行再建術を行った群は統制群に比べより若く，男性が多く，糖尿病や心臓病の有病率が低く，血行再建術や再梗塞の既往歴が多く，処方されている薬の種類が多かった．様々な共変量を調整した結果，血行再建術により死亡率が下がることがわかり，急性心筋梗塞後の早期侵襲的手術を支持する結果となった．

(7) Ayanian ら，[5] の心臓外科医の診察についての研究

急性心筋梗塞の予後は，退院後外来で受けるケアによって影響を受ける可能性がある．特に心臓外科医による診療の効果は先行研究の知見は一貫していなかった．

そこで，Ayanianらはアメリカの7州35,520人（全員が65歳以上）のデータを集め，退院後3ヶ月以内に心臓内科医（cardiologist）の診療を受けた群と，内科医（internist）もしくは家庭医（family practitioner）の診療は受けたが専門医の診療は受けていない群とで2年後の死亡率を調べた．心臓内科専門医の診療を受けていない群は，相対的に若く，白人が多く，男性が多く，並存症状が少なく，入院中侵襲的治療を受けた割合が多かった，そこで傾向スコアを使ってこれらの変数を調整した結果，心臓内科医にかかった群の方が死亡率は低いことがわかった．

(8) MacKenzie ら，[47] による外傷センター（Trauma Center）の有用性の研究

MacKenzieらは米国の14州において，レベル1の外傷センターのある病院（n=18）と，外傷センターを持たない病院（n=51）とで治療を受けた患者の死亡転帰を比較した．原データでは外傷センターの無い病院で治療を受けた患者群は相対的に年齢が高く，より並存症状があり，女性や白人・保険加入者が多く，症状の程度は軽かった．傾向スコアによってこうした共変量調整した結果，外傷センターを持つ病院に入院中の死亡率は，持たない病院に比して有意に低く（7.6% vs. 9.5%），また1年以内の死亡率も有意に低い（10.4% vs. 13.8%）ことが確かめられた．

このように，臨床医学・疫学・医療経済など様々な分野において傾向スコアを用いた解析が行われていることがわかる．

他にも欧米の一流誌において傾向スコアを用いた解析は非常に多数報告されているが，特にNew England Journal of Medicine(NEJM)，Journal of the American Medical Association(JAMA)，Lancetに掲載された応用研究の一部について（引用文献[48]－[55]）は，表1に

- サンプルサイズ（"N"）
- 従属変数
- 独立変数
- 共変量の数
- 具体的な共変量（デモグラフィック，患者の他の病気／病歴，他に投与された薬，他に受けた手術／検査／治療，病院の性質）
- 傾向スコアの利用の仕方（"PS方法"）
- 共変量の選択基準
- 強く無視できる割付（Strong Ignorability）の仮定のチェックの確認方法
- 調整前の結果と後の結果（オッズ比など）
- （あれば）他の解析法をしているか？
- 先行研究との知見の一貫性

といった観点に分けてまとめた．

9. いくつかの拡張について

（1）条件が３つ以上の比較への拡張

これまで紹介してきた傾向スコア解析は全て２群での因果効果の推定についてのものである．同時に解析の対象にする集団が３つ以上の場合，Rubin,[35] は２群ごとに比較をすることを薦めている．しかしこれを行うと，各解析で母集団とするものが異なるという問題が生じる．例えばA，B，Cの3群が存在する場合，群A，Bの傾向スコアによる解析はAとBのそれぞれの母集団の（しかもそれぞれのサンプルサイズの比の）混合母集団についての推測をすることになり，A，B，C全体を母集団とした場合の結果を与えることができない．これに対してImbens,[56] は，たとえ割付変数が二値でなくても，傾向スコアを利用できることを証明した（多群での傾向スコアは一般化傾向スコア(generalized propensity score)と呼ばれる）．その方法は非常に簡単であり，例えば３群ならば３カテゴリーの名義ロジスティック回帰モデルの各群への所属の予測確率を一般化傾向スコアとし，その逆数を重み付けとして群Aの周辺平均を求めればよいというものである．またはAについての一般化傾向スコアを「群Aに所属するか，群A以外に所属するか」を２値の従属変数としたロジスティック回帰を用いて傾向スコアを算出してもよい．

（2）Doubly　Robust 推定

また，傾向スコアを用いた調整法では「割付を共変量を用いて説明するモデル」を誤って設定すれば推定に偏りが生じるということが知られている．そこで，「従属変数を共変量によって説明するモデル＝共分散分析的モデル」と「割付を共変量によって説明するモデル＝傾向スコアを推定するためのモデル」どちらも用いて因果効果を推定するDoubly Robust 推定(Rotnitzky & Robins,[57], Bang & Robins, [58], Hoshino[59] が提案されており，この方法はどちらかのモデルが正しければ，因果効果の推定を正しく行うことでき，かつ検定の検出力も高いことが知られている．この手法は繰り返し計算を含む複雑な手法であるためにまだ実際の解析例では利用されていないが，今後プログラムが整備されれば利用されていく可能性が高いと考えられる．

（3）傾向スコア較正（Propensity Score Calibration）

これまでも述べたように傾向スコアを用いた共変量調整においては，先行研究においてすでに取り上げられ，かつ従属変数と割付それぞれに影響を与えていると考えられる共変量を利用する必要がある．しかしすでに得られているデータや政府などが行っている大規模調査などの二次データを本調査（main study）のために利用する際には，必要な共変量が測定されていない場合がある．このような場合において，必要な共変量すべてを測定した別の調査(validation study)を行い，そこで推定することができる「本調査において測定されている共変量」と「必要な共変量」との回帰関係を利用して，本調査での調整に利用する傾向スコア較正(Propensity Score Calibration）がStürmer, Schneeweiss, Avorn & Glynn,[60] によって提案されている．これはこれまでも医学研究で利用されている回帰の較正（Regression Calibration, Rosner, Willett & Spiegelman,[61], Fraser & Stram,[62]）の一種であるが，本調査で測定されていない共変量が多次元の場合でも利用が可能であるという点でより有用である．

10．まとめ

本稿で紹介した傾向スコアを用いた調整法は様々な応用研究に利用されており，すでにNew England Journal of MedicineやLancetなどの欧米のトップジャーナルにもその応用例が数多く掲載されているなど，統計手法として一般的なものになりつつある．この手法の最大の利点は，共分散分析的な手法と異なり，複数の共変量と従属変数の回帰関係を特定しなくても因果効果を推定できるところにあり，統計学的には仮定の少ない“セミパラメトリックな”頑健な手法であることが様々な理論的検討によって示されている．

但し，調整にあたっては，何を共変量として利用するかが非常に重要である．これは「強く無視できる割り当て」条件の成否に関係する．しかし応用例の紹介で見たように，これまでの応用研究では「強く無視できる割り当て」条件の成立をチェックしているものはあまりない．少数の研究で，ロジスティック回帰のフィットが報告されるのみである．

多くの応用研究においては，先行研究や理論上，従属変数に関連があると考えられる変数が共変量として利用されている．

共変量選択の問題は依然存在するが，「強く無視できる割り当て」条件が成立するように共変量を十全に選択することは実際には難しい．しかし完全な共変量のセットを探索することに固執するよりは，現時点のデータで利用可能，かつ理論的に考慮に値する共変量を用いて調整を行うことで，当該分野の研究が漸進することにこそ共変量調整を行う意義があると考えられる．つまり「先行研究では…という共変量を考慮にいれて傾向スコアを用いて解析した．結果…であった．今回はさらに…を共変量として考慮して解析した．その結果…」といったように，調整する意味のある共変量が徐々に同定されていくことこそが重要であろう．このことは傾向スコアを用いる／用いないに関わらず，これまでも実証科学のあらゆる分野で行われてきた研究の流れであり，科学的な知見をより安定したものとするプロセスとして重要である．

また，９節においても述べたように，周辺期待値を求める際には，母集団は何かを明確にすることが応用研究において非常に重要である．例えば実験群が9000人，対照群が1000人のデータの場合，傾向スコアを用いた調整後の因果効果の推定値は「実験群：対照群」を9：1で混合した母集団における期待値の推定値になってしまう．従って，逆に１：９の比で混合した場合の結果と大きく異なる可能性があることに注意するべきである．このように，適用の際にはどの集団を目標として調整がされるのかに対して十分注意する必要があるが，このことは欧米での応用研究でも未だほとんど触れられていない（この問題を回避するには，例えば星野，[32] などを参照）．近年より応用研究に即した形での理論的な検討が数多く行われてきており，今後このような観点に関しても注意が必要となる可能性が高い．

引用文献

1) 丹後俊郎．良質の根拠を生む randomization の本質―科学研究者としてのセンス―. 公衆衛生研究 2000；49：308-312.

2) Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.

3) Gum PA, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease, JAMA 2001;286:1187-1194.

4) Reinisch JM, Sanders SA, Mortensen EL, Rubin DB. In utero exposure to phenobarbital and intelligence deficits in adult men. JAMA 1995;274: 1518-1525.

5) Ayanian JZ, Landrum MB, Guadagnoli E, Gaccione P. Specialty of ambulatory care physicians and mortality among elderly patients after myocardial infraction. New England Journal of Medicine 2002;347: 1678-1686.

6) Wang PS, Schneeweisse S, Avorn J, Fischer MA, Mogun H, Solomon DH, Brookhart MA. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. New England Journal of Medicine 2005;353: 2335-2341.

7) Petersen LA, Normand ST, Daley J, McNeil BJ. Outcome of myocardial infarction in veterans health administration patients as compared with medicare patients. New England Journal of Medicine 2000;343: 1934-1941.

8) Coyte PC, Young W, Croxford R. Costs and outcomes associated with alternative discharge strategies following joint replacement surgery: Analysis of an observational study using a propensity score. Journal of Health Economics 2000;19: 907-929.

9) Hill J, Waldfogel J, Brooks-Gunn J. Differential effects of high quality child care. Journal of Policy Analysis and Management 2002;21: 601-627.

10) Lu B, Zanutto E, Hornik R, Rosenbaum PR. Matching with doses in an observational study of a media campaign against drug abuse. Journal of the American Statistical Association 2001;96:1245-1253.

11) Newman JH, Pradham M, Rawlingsm LB, Ridder G, Coa R, Evia JL. An impact evaluation of education, health, and water supply investments by the Bolivian Social Investment Fund. The World Bank Economic Review 2002;16:241-274.

12) Bingenheimer JB, Brennan RT, Earls FJ. Firearm violence exposure and serious violent behavior. Science 2005;308:1323-1326.

13) 佐藤俊哉．傾向スコアを用いた因果効果の推定．柳井晴夫，岡太彬訓，繁桝算男，高木廣文，岩崎学，編．多変量解析実例ハンドブック．東京：朝倉書店；2002. p. 240-250.

14) Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 1974;66: 688-701.

15) Neyman JS. On the application of probability theory to agricultural experiments. essay on principles. section 9. (Tlanslated and edited by Dabrowska DM. Speed TP. Statistical Science 1990;5:465-480) Annals of Agricultural Sciences 1923;10:1-51.

16) 星野崇宏，繁桝算男．傾向スコア解析法による因果効果の推定と調査データの調整について．行動計量学 2004；31:43-61.

17) 狩野裕．構造方程式モデリング，因果推論，そして非正規性．甘利俊一，狩野裕，佐藤俊哉，松山裕，竹内啓，石黒真木夫，編．多変量解析の展開．東京：岩波書店；2002. p. 64-130.

18) Cook TD, Campbell DT. Quasi-experimentation : design & analysis issues for field settings. Boston: Houghton Mifflin;1979.

19) Little RJA, Rubin DB. Statistical analysis with missing data. New York:Wiley;1987.

20) Rubin DB. Inference and missing data. Biometrika 1976;63: 581-590.

21) Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 1965;58: 295-300.

22) Rosenbaum PR, Rubin DB. (1985). The bias due to incomplete matchings. Biometrics 1985;41:103-116.

23) 佐藤俊哉 , 松山裕. 疫学・臨床研究における因果推論. 甘利俊一 , 狩野裕 , 佐藤俊哉 , 松山裕 , 竹内啓 , 石黒真木夫 , 編. 多変量解析の展開. 東京 : 岩波書店 ;2002. p.131-176.

24) Rosenbaum PR. (2002). Observational studies. 2nd edition. New York: Springer-Verlag; 2002.

25) Rosenbaum PR, Rubin DB. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 1984;79:516-524.

26) Rosenbaum PR, Rubin DB. (1985). The bias due to incomplete matchings. Biometrics 1985;41:103-116.

27) Rubin DB. The use of propensity scores in applied Bayesian inference. In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM, editors. Bayesian Statistics 2.North-Holland: Elsevier Science Publisher B.V.,; 1985. p. 463-472

28) Rosenbaum PR. Model-based direct adjustment. Journal of the American Statistical Association 1987;82:387-394.

29) Horvitz D, Thompson D. A generalization of sampling without replacement from a finite population. Journal of the American Statistical Association 1952;47:663-685.

30) Rotniztky A, Robins JM. Semiparametric regression estimation in the presence of dependent censoring. Biometrika 1995;82:805-820.

31) Liang K-Y, Zeger SL. (1986). Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13-22.

32) 星野崇宏 , 欠測群の周辺分布の母数に対する傾向スコアを用いた重み付き M 推定量の提案と介入効果研究への応用. 行動計量学 2005;32: 121-132.

33) Hoshino T, Kurata H, Shigemasu K. A propensity score adjustment for multiple group structural equation modeling. Psychometrika. 2007. http://www.springerlink.com/content/1860-0980/?sortorder=asc&Content+Status=Accepted

34) Huber PJ, Robust statistics. New York: John Wiley; 1981.

35) Rubin DB. Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine 1997;127:757-763.

36) Drake C. Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. Biometrics 1993;49:1231-1236.

37) Cepeda MS, Boston R, Farrar JT, Storm BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. American Journal of Epidemiology 2003;158:280-287.

38) D'Agostino Jr, RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine 1998;17:2265-2281.

39) Rosenbaum PR, Rubin DB. Assesing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society, Series B, 1983;45: 212-218.

40) Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiology and Drug Safety 2004;13:841-853.

41) 星野崇宏 , 前田忠彦. 傾向スコアを用いた補正法の有意抽出による標本調査への応用と共変量の選択法の提案. 統計数理 2006;54:191-206.

42) Brookhart MA, Schneeweiss S, Rothmanm KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. American Journal of Epidemiology 2006;163:1149-1156.

43) Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment, Journal of the Royal Statistical Society, Series A, 1984;147:656-666.

44) McWilliams JM, Zaslavsky AM, Meara E. Ayanian JZ. Impact of medicare coverage on basic clinical services for previously uninsured adults. JAMA 2003;290:757-764.

45) Shishehbor MH, Litaker D, Pothier CE, Lauer MS. Association of socioeconomic status with functional

capacity, heart rate recovery, and all-cause mortality. JAMA 2006;295:784-792.

46) Stenestrand U, Wallentin L. Early revascularisation and 1-year survival in 14-day survivors of acute myocardial infarction: a prospective cohort study. Lancet 2002;359:1805-11.

47) MacKenzie EJ, Rivara FP, Jurkovich GJ, Nathens AB, Frey KP, Egleston BL, Salkever DS, Scharfstein DO. A national evaluation of the effect of trauma-center care on mortality. New England Journal of Medicine 2006;354:366-378.

48) Abidov A, Rozanski A, Hachamovitch R, Hayes SW, Aboul-Enein F, Cohen I, Friedman JD, Germano G, Berman DS. Prognostic significance of dyspnea in patients referred for cardiac stress testing. New England Journal of Medicine 2006;353:1889-98.

49) Hannan EL, Racz MJ, Walford G, Jones RH, Ryan TJ, Bennett E, Culliford AT, Isom OW, Gold JP, Rose EA. Long-term outcomes of coronary-artery bypass grafting versus stent implantation. New England Journal of Medicine 2005;352:2174-83.

50) Lindenauer PK, Pekow P, Wang K, Mamidi DK, Gutierrez B, Benjamin EM. Perioperative beta-blocker therapy and mortality after major noncardiac surgery. New England Journal of Medicine 2005;353:349-61,

51) Mangano DT, Tudor IC, Dietzel C. The risk associated with aprotinin in cardiac surgery. New England Journal of Medicine 2006;354:353-65.

52) Mehta RL, Pascual MT, Soroko S, Chertow GM. Diuretics, mortality, and nonrecovery of renal function in acute renal failure. JAMA 2002;288:2547-53.

53) Schneeweiss S, Walker AM, Glynn RJ, Maclure M, Dormuth C, Soumerai SB. Outcomes of reference pricing for angiotensin-converting-enzyme inhibitors. New England Journal of Medicine 2002;346:822-9.

54) Vikram HR, Buenconsejo J, Hasbun R, Quagliarello VJ. Impact of valve surgery on 6-month mortality in adults with complicated left-sided native valve endocarditis: A propensity analysis. JAMA 2003;290:3207-3214.

55) Welch RD, Zalenski RJ, Frederick PD, Malmgren JA, Compton S, Grzybowski M, Thomas S, Kowalenko T, Every NR. Prognostic value of a normal or nonspecific initial electrocardiogram in acute myocardial infarction. JAMA 2001;286:1977-84.

56) Imbens GW, The role of the propensity score in estimating dose-response functions. Biometrika 2000;87:706-710.

57) Rotniztky A, Robins JM. Analysis of semi-parametric regression models with nonignorable non-response. Statistics in Medicine 1997;16:81-102.

58) Bang H, Robins JM. Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics 2005;61:962-972.

59) Hoshino T, Doubly robust type estimation for covariate adjustment in latent valiable modeling. Psychometrika. (in press). http://www. springerlink. com/content/1860-0980/sortorder=asc & Content+ status=Accepted

60) Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting Effect Estimates for Unmeasured Confounding with Validation Data using Propensity Score Calibration. American Journal of Epidemiology 2005;162:279-289.

61) Rosner B, Willett WC, Spiegelman D. Correction of Logistic Regression relative risk estimates and confidence intervals for systematic within-person measurement error. Statistics in Medicine 1989;8:1051-1069.

62) Fraser GE, Stram DO. Regression calibration in studies with correlated variables measured with error. American Journal of Epidemiology 2001;154:836-844.

表1：一流紙に掲載された傾向スコアを用いた応用研究の一部

筆頭著者 出版年雑誌	N	従属変数	独立変数	共変量	共変量					
					デモグラフィック	病気・病歴	薬	手術·検査・治療	病院	その他
Abidov (2005) NEJM 48)	1,968	死亡率	呼吸困難	15	年齢，性別	糖尿病，高血圧，高コレステロール血症，左室肥大 左室拡張，心房細動		心筋血流 SPECT，喫煙，Q 波，心拍数，ストレス，ストレス性虚血		
Ayanian (2002) NEJM 5)	20,398	死亡率	心臓外科医への外来	36	年齢，性別，人種・民族，州	心筋梗塞，狭心症，心不全，脳卒中，末梢血管疾患，高血圧，糖尿病，慢性閉塞性肺疾患，運動障害，認知症，入院中の合併症	アスピリン，β阻害薬，ACE 阻害薬，コレステロール低下薬	循環器医のコンサルティング，心エコー図，ストレス，冠動脈造影，冠動脈形成術，冠動脈バイパス手術	田舎か，病院教育，経営，冠動脈造影設備，バイパス手術設備，介護ケア施設，心臓手術設備	
Gum (2001) JAMA 3)	2,702	死亡率	アスピリン使用	34	年齢，性別	糖尿病，高血圧，冠動脈疾患，冠動脈大動脈バイパス移植術，経皮的冠動脈形成術，非 Q 波心筋梗塞，心房細動，心不全，喫煙，肥満	ジゴキシン，β阻害薬，ベラパミル・デイルチアゼム，ニフェジピン	血清脂質降下療法，心臓血管系状態，BMI，心駆出率，心拍数，血圧，腰痛，Mayo リスク指標，最大運動耐容尺度，運動後心拍数，虚血性 ECG，左心室駆出率，虚血性心エコー図		
Hannan (2005) NEJM 49)	59,314	死亡率	冠動脈バイパス術 vs 経皮的冠動脈形成術	22	性別，人種・民族	心筋梗塞，糖尿病，慢性閉塞性肺疾患，頸動脈病変，大腿・膝窩疾患，腎不全（要透析）		左室駆出分画		
Lindenauer (2005) NEJM 50)	335,922	死亡率	β阻害薬投与	−	年齢，性別，人種・民族，保険	高血圧，糖尿病，虚血性心疾患，腎不全，高脂血症，脳血管疾患	抗生剤，深部静脈血栓症予防薬，脂質低下薬，Ca チャネル阻害薬，ACE 阻害薬，抗凝固薬，ループ利尿薬，アンギオテンシン受容体阻害薬，チアジド，抗不整脈薬，ドーパミン・ドバトミン	手術タイプ（血管，整形外科，腹部，胸部，その他），入院の緊急度，入院日数，医療費，RCRI 得点，拡張 RCRI 得点，手術リスク	田舎か，病院教育，ベッド数，	交互作用
MacKenzie (2006) NEJM 47)	5,191	死亡率	外傷センター	−	年齢，性別，人種・民族，保険	凝固障害，肥満		Charlson 同時罹患得点	地域	（基準不明）
Mangano (2006) NEJM 51)	4,035	器官障害	アプロチニン使用	45	性別，人種，教育	糖尿病，高血圧，狭心症，心不全，心筋梗塞，心ブロック，頚動脈疾患，脳卒中，肝臓病，腎臓病，肺疾患，弁疾患		冠動脈バイパス手術，弁手術，経皮経管冠動脈形成術，大動脈血管手術，緊急手術，冠動脈ステント，冠動脈アテレクトミー，心駆出率，クレアチン		交互作用
McWilliams (2003) NEJM 44)	2,203	各種検査受診	保険加入	12	年齢，性別，居住地域，教育，雇用，年収	健康状態（自己申告），アルコール摂取，喫煙，慢性疾患				
Mehta (2002) JAMA 52)	552	死亡率	利尿剤	5	年齢	腎毒性病因（腎不全の），急性呼吸不全，心不全		血清尿素窒素		
Schneeweiss (2002) NEJM 53)	37,362	病院使用	新 ACE 阻害薬への変更	−	年齢，性別，年収，保険，支出	緊急救命室回数，来院回数，異なる診断回数				
Shishehbor (2006) JAMA 45)	7,158	死亡率	SES	36	年齢，性別，人種・民族，居住地域（収入，学歴，管理職，不動産価値等），雇用，病院までの距離，保険	糖尿病，高血圧，喫煙，心臓病家族歴，冠動脈疾患，心筋梗塞，経皮的冠動脈形成術，末梢血管疾患，脳血管障害		BMI	慢性疾患スコア	交互作用（有意）
Stenestrand (2002) Lancet 46)	1,835	死亡率	血行再建術	33	年齢，性別	喫煙，心筋梗塞，経皮的冠動脈形成術，冠動脈バイパス術，糖尿病，高血圧，循環停止	アンギオテンシン変換酵素阻害薬，抗凝固薬，アスピリン，β阻害薬，カルシウムチャネル阻害薬，ジギタリス，利尿薬，持続性硝酸塩，スタチン	ECG，血栓溶解療法，静脈β阻害薬，抗凝固薬，ニトログリセリン，心房細動，心不全，再梗塞，心エコー図，入院年		
Vikram (2003) JAMA 54)	218	死亡率	弁手術	−	年齢，性別，人種·民族，気温，	免疫無防備状態，HIV，AIDS，ドラッグ使用，心不全，精神状態，黄色ブドウ球菌，緑色レンサ球菌，血清クレアチン，塞栓，真菌血症，難治性感染症		心エコー図，心駆出率，Charlson 同時罹患得点 (Charlson 1987)	年間心筋梗塞患者数，病院教育，心カテーテル設備，緊急血行再建術割合	
Welch (2001) JAMA 55)	～27,384	死亡率	急性心筋梗塞患者の ECG	21	年齢，性別，人種・民族	高コレステロール血症，急性心筋梗塞，経皮的冠動脈形成術，冠動脈バイパスグラフト術，心不全 (Killip 分類Ⅱ)，喫煙	アスピリン，ニトログリセリン阻害薬，リドカイン，ACE 阻害薬，Ca チャネル阻害薬	梗塞位置，心エコー図，再灌流療法，抗血小板療法，心臓カテーテル	病院固有番号	

PS 方法	共変量基準	Strong Ignorability	調整前	調整後	他の手法	知見一貫性
マッチング	全変数 (nonparsi-monious)	–	1.90 (HR)	1.76 (HR)	–	関連は指摘されていたが，体系だった疫学研究はなかった
マッチング	先行研究	共変量バランス (>90％)	0.57 (OR)	0.76 (OR)	–	先行研究の知見は一貫していなかった
マッチング	全変数 (nonparsi-monious)	c=0.83	1.00 (OR) 4.5% vs 4.5%	0.54 (OR) 4% vs 8%	傾向スコア＋共変量の同時調整	罹患率や短期死亡率への影響を示した研究はあったが，長期死亡率への影響を示したのは初めて
層化	ロジスティック回帰で有意だった変数	–	1.05 (HR)	0.76 (HR)	–	比較研究はあるが，技術革新前のものであった
マッチング	全変数 (nonparsi-monious)	–	1.15 (OR)	0.99 (OR)	心リスクで層化した所，高リスク群では死亡率を下げることが判明	先行研究では限られた知見しかなかった
重み付き推定	全変数 (nonparsi-monious)	共変量バランス，トリミング	1.39 (OR) 8.0% vs 5.9%	0.78 (OR) 7.6% vs 9.5%	–	先行研究は研究デザインや信頼性に問題があった（特にバイアス調整）
共分散分析	全変数 (nonparsi-monious)	c=0.71	3.18 (OR)	2.34 (OR)	–	懸念はあったが直接リスクを示す研究がなく，アプロチニンは使用されてきた
–	–	共変量バランス	−15.3 (CP)	−11.8 (CP)	–	結果は予想されるものであり，確かに説明力があることが示された
共分散分析	有意差があった変数 (p<.25)	–	1.37 (OR)	1.68 (OR)	(単なる) 共分散分析：OR1.65	目だった先行研究はなかった
共分散分析	–	c=0.60	1.03 (RaR)	1.01 (RaR)	–	他国での類似研究はあったが，制度・文化などが異なり比較不可能
マッチング	先行研究	c=0.85	1.51 (HR)	1.84 (HR)	–	SES が死亡リスクなどを高めることは知られていた
共分散分析，層化	死亡率を目的変数とする Cox 回帰で有意だったもの	C=0.87, Hosmer-Lemeshow 検定 (p=0_007).	0.44 (RiR)	0.53 (RiR)	–	先行研究では議論が分かれていた
マッチング	全変数 (nonparsi-monious)	c=0.86，回帰診断（残差，はずれ値，多重共線など）	0.43 (HR)	0.45 (HR)	–	有用性を示した研究は少なく，共変量の影響を考慮した疫学研究は初めて
マッチング，共分散分析	「死亡率と関連する (relevant) 変数」	c=0.82，共変量バランス (95%)	0.47 (OR)	0.58 (OR)	傾向スコア＋共変量の調整：OR0.59	関連は指摘されていたが，ECG 単独の予測力を示す研究はなかった

注 OR：Odds Ratio（オッズ比），RaR：Rate Ratio（率比）HR：Hazard Ratio（ハザード比），RiR：Risk Ratio（リスク比），CP：Changes in Percent(% 変化）