# Propensity Scores

## ランダム化臨床試験でなくとも N Engl J Med に掲載される

β -blocker を術前に用いた方がよいとするランダム化エビデンスが無いにもかかわらず、その有効性が一般的に信じられている。この研究では後ろ向きに、β -blocker の有効性を propensity score というものを用いて検討していた。そして low risk では関係ないが、high risk 患者さんにβ -blocker を用いると術後死亡リスクを減少させることができると結論している。
(Lindenauer PK, Pekow P, Wang K, Mamidi DK, Gutierrez B, Benjamin EM. Perioperative beta-blocker therapy and mortality after major noncardiac surgery. N Engl J Med. 2005 Jul 28;353(4):349-61.)。

## propensity score を用いる背景

臨床の世界において、ある治療方針を決定するには実に多くの要素から判断する。しかし一般的には悪いからより強い治療をという傾向があることは否めない。そして、その治療成績が悪くても患者さんの状態が悪かったから、予後不良因子を多く兼ね備えていたからしょうがない、という論点に帰結してしまう。しかし本当にそうだろうか？これを明らかにするにはランダム化臨床試験が最も信頼性の高いデータを出してくれる。しかし、既に一般的になってしまった治療において、１人１人の患者さんにベストを尽くそうとしている医師が、自分の経験からくる治療方針を棚上げしてランダム化に賛同するだろうか？また、既に保険で認められている薬の用法に対して、再度ランダム化臨床試験を行う製薬会社があるだろうか？

## propensity score の概念

最近 propensity score という方法が考案された。観察的データのみからランダム化臨床試験に近い結果を得る手法として注目されている。Propensity とは性癖、性質、傾向を意味する。ある治療を行うか否かは、医師の性癖（好み）に寄る部分が大きい。ある患者さんが A 医師にかかった場合には治療 A が選択され、同じ患者さんが B 医師にかかれば、治療 A は選択されずに経過観察されるだけかもしれない。そこで、多くの因子のデータを基に治療 A が選択される確率を多ロジスティック解析を用いて算出する。これを propensity score と呼ぶ。これは確率であるため 0 から 1 の間にあり、2 人の

患者さんの propensity score が 0.6 であれば同じ確率（およそ 3 人に 2 人）でその治療を受けることになる。そこで、propensity score が同じ 0.6 であったにもかかわらず、ある患者さんは治療 A を受け、ある患者さんは受けなかったという状況が発生する、そこで、このような 2 人をマッチさせる。そうすると、誰がみても治療 A が選択される場合（患者さん）、誰がみても治療 A が選択されない場合（患者さん）は propensity score を用いた解析からは除かれることになる。

propensity score でマッチングすると、自然と予後因子が治療 A を行わなかった患者群と行った患者群で一致してくる。この表はあたかもランダム化臨床試験で最初にでてくる表のようである。すなわち、propensity score を用いた研究は観察であるにもかかわらずランダム化臨床試験のようになるのである。実際の臨床現場において医師は propensity score が高くてもその治療を行なわなかったり、逆に低い場合でもその治療を行なったりする。つまり中間の部分では同じ予後因子を持っていても、ある人はその治療を受け、またある人は治療を受けない。これはまさにランダム化臨床試験と似ているわけで、同じ背景因子をもつ患者さんでも医師によって治療方針が異なることを逆手にとったのが propensity score を用いた臨床観察研究ということになる。これを、一言でいうならば、治療選択バイアスを軽減するためのデザインである。

しかし、いくつか問題点も存在する。１つは治療群間でマッチングできる対象が少ない場合である。例えば、ある疾患に対して治療 A が選択される基準は、医師の裁量に寄らずほとんど一定しているときなどがそれに相当する、何故なら、マッチングできる対象が、例えば全体の数%にまで低下してしまうからである。第二に、変数が少ない場合である。なるべくランダム化臨床試験に近づけたいのであるが、十分な変数がとれていない場合には、通常の多変量解析で補正しても問題ない（多変量解析より propensity score の方が有用である場合については後半に解説する）。

1996 年 Conners らは ICU に入院した患者を対象に右心カテーテルの予後に与える影響を propensity score を用いて研究し JAMA に発表した。本論文が propensity score のはしりでもあるので、以下この論文を紹介することにする。

The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients

Connors AF, et al. JAMA 276:889-97,1996.

要旨

背景：ICU に入院して 24 時間以内に右心カテーテル（RHC）を行なった患者さんで、生存期間、入院期間、ケアの強さ、治療費を調べた。
デザイン：prospective cohort study
Setting: 1989-1994 年、5 つのアメリカ教育病院
対象：９つの予め決めておいた疾患カテゴリーの１つのために ICU に入院してケアを受けた重症疾患の成人患者さん 5735 人。
主な結果の測定方法:生存期間、治療費、ケアの強さ、在院日数、はカルテおよび National Death Index より検索した。RHC の propensity score は multivariate logistic regression を用いて創り出した。Case-matching と multivariable regression modeling は propensity score を用いて RHC を用いたものと用いなかったものとの間で調整を行なった後、特別な結果と RHC の関連を調べるのに用いられた。Sensitivity analysis は結果に含まれる認識されなかった因子の可能性のある効果を評価するのにもちいられた。
結果：case-matching analysis では、RHC を受けた患者さんでは day 30 の死亡率が高くなっていた(OR 1.24, 95% CI 1.03 – 1.49)。RHC を受けた患者さんの入院 1 日当りの費用は$49,300 ($17,000; 25th, $30,500; 50th, 56,600;75th)であり、RHC を受けなかった患者さんの入院 1 日当りの費用は$35,700 ($11,300; 25th, $20,600; 50th, 39,200;75th)であった。RHC を受けた患者さんの ICU の平均滞在期間は 14.8 日(5, 9, 17)、RHC を受けなかった患者さんの ICU の平均滞在期間は 13.0 日(4, 7, 14)であった。これらの所見は全て multivariate modeling techniques により確認された。Subgroup analysis では RHC と予後との関係を左右するような特別な subgroup を同定することはできなかった。入院 2 ヶ月後最も死亡率を上げる危険因子は RHC であった。
結論：selection bias を調整した後の重傷患者を対象にした本観察的研究では、RHC

は死亡率の増加と資源の過剰利用に関係していた。何故 RHC の利点がないのに頻繁に行なわれているかは不明である。この解析結果は他の観察的研究で確認されるべきである。これらの所見は RHC のランダム化臨床試験を行なうに値するとおもわれた。

以下の点に注意して論文を読み進めていくことにする。

1．　この解析に含まれた患者さんは単にSUPPORTに含まれた患者さん全員の一部というだけだろうか？

2．　この研究における介入，結果，交絡について説明せよ。

3．　7人の専門家がRHCの可能性のある予測因子を同定しているが、本当はRHC予測因子であるのに選択された予測因子から漏れてしまったものがあった場合、propensity scoreに含まれるだろうか？

4．　何人の患者さんがこのmatched analysisに組み入れられたのか？何故減ってしまったのか？

5．　ICU入院後24時間以上経ってからRHCを行なった患者さんは主な解析から外された。もし含めるとしたらどちらのグループに入れるべきだろうか？晩期RHCの二次的解析というのはどうだろうか？

6．　RHCを行なった患者さんのpropensity scoreの平均は0.577であり、95%信頼区間は0.108 – 0.943であり、行なわなかった患者さんのそれは0.253 (95% 信頼区間 0.011 – 0.779) と記載されているが、正しい結果だと思うか？

7．　Propensity scoreに含めた予測因子のいくつかに対して、RHC施行群と非施行群に分けて再度差がないかどうか検討しているが(Table 3)、なぜこのようなことをする必要があるのか？

8．　RHCを施行する病院は施行しない病院より死亡率が高いことになる。どのような状況下においてRHCと患者生存の関係をバイアスなく述べることができると思いうか？特に、RHCと基本的予後因子は比較された病院間でお互い独立している必要はあるだろうか？discussionにおいて、著者らはRHCが単にケアの強い治療を好む医療のマーカーである可能性を示唆している。

## 研究の背景

右心カテ(RHC)により肺動脈の情報を得ることは重症患者にとって重要であり、患者さんにより良い結果をもたらすと考えられている。ランダム化臨床試験によって RHC の利点はまだ証明されていないが、既に広く普及し利点が大きいと信じられているため、ランダム化臨床試験を施行するのは困難である。最近ランダム化臨床試験を施行しようとしたが、多くの医師がランダム化に反対し中止となっています。そのため過去観察的研究が行なわれてきたのが実情である。特に老人と心筋梗塞の患者さんで RHC を行うことにより死亡率が高くなるとする報告もある。しかし、RHC の適応は医師の裁量に委ねられ、例えば血圧の低い患者さんに RHC が行なわれやすい傾向があるとすると、死亡率が高くなって然るべきである。このような選択バイアスは confounding by indication と呼ばれています。Rosenbaum & Rubin らによって提案された propensity score はこのような選択バイアスを調整するのには強力である。Propensity score は、その行為を決定する全ての因子を多ロジスティック解析によって算出される。この propensity score を用いてマッチングすることによって対象となる患者さんは RHC を等しい確率で施行されることになる。

## 研究方法の詳細

### 対象

The Study of Understand Prognosis and Preferences for Outcomes and Risks of Treatments (SUPPORT) は５つのセンターを介して入院成人重症患者に対する結果評価と decision making について調査したものである。以下の５つ医療センターが参加した。Beth Israel Hospital (Boston, MA), Duke University Medical Center (Durham, NC), Metro Health Medical Center (Cleaveland, Ohio), St Joseph's Hospital (Marshfield, Wis), University of California Medical Center (LA)。研究は George Washington 大学がまとめた。統計解析は Duke 大学が担当した。

SUPPORT で対象としたのは病院あるいは ICU に入院し６ヶ月後の死亡率が 50%と予想される重症疾患分類９つのうちの１つ以上に当てはまる患者さん達である。疾患カテゴリーは、急性呼吸不全(acute respiratory failure: ARF), 慢性閉塞性肺疾患(chronic

obstructive pulmonary disease: COPD), うっ血性心不全 (congestive heart failure: CHF), 肝硬変(cirrhosis)、非外傷性昏睡(nontraumatic coma)、大腸癌肝転移、非小細胞型肺癌、癌あるいは敗血症に伴う多臓器不全である。除外基準は 18 歳未満、48 時間以内の死亡ないし退院、英語をしゃべれない、急性精神疾患、妊娠、AIDS、急性熱傷、頭部外傷ないしは他の外傷で急性呼吸不全あるいは多臓器不全を伴わないものであった。全ての患者さんは 6 ヶ月まで経過観察された。この計画の詳細は過去に発表されている。

この研究では 24 時間以内に ICU に入院となった 3735 人の SUPPORT 患者さんを対象とした。RHC を行なったかどうかは、カルテの記載と看護記録より判断した。Exposure は 24 時間以内に行なわれた RHC とし、2184 人の患者さんが対象となった。

各病院の前日入院の患者さん全てをスクリーニングし、ICU では毎日のミーティングにおいて SUPPORT の条件に当てはまるかどうか検討してもらった。もしも患者さんが条件を満たしていれば、詳細にカルテを検討した。研究プロトコールは各疾患の診断を確認するとともに重症度をも評価した。入院時の診断が複数ある場合には４つまでとした。研究に含まれる前の入院期間も調べてある。生理的状態およびケアの強さは入院中 1, 3, 7, 14, 25 日の時点で検討している。生理的状態は入院後 24 時間の値；体温、平均血圧、心拍数、呼吸数、PaO2, FIO2, PaCO2, pH, Na, K, Ht, WBC, アルブミン、ビリルビン、クレアチニン、Glasgow Coma Score の異常をもって判断している。抜けている値は正常のものとして判断した。Acute Physiology and Chronic Health Evaluation (APCHE)III score が計算された。2 ヶ月の時点の予後は SUPPORT で開発した統計にて評価した。ケアの強さは Therapeutic Intervention Scoring System (TISS)で評価した。しかし RHC に直接関係する TISS のポイントは除外してある。

インタビュー

患者さん、あるいは患者さん代理人に対してよく訓練を受けたインタビュー者が 3 日後（2-6 日後）にインタビューを行なっている。インタビューでは年齢、性、人種、教育、収入、保険について聞いている。もしも患者さんの意識がない、挿管中、インタビューに受け答えできないような場合には代理人から情報を聞き出している。もしも両方から情報が得られなかった場合にはカルテから情報を引き出している。生活機能に関しては Katz Activities of Daily Living Scale (ADL) および Duke Activity Status Index (DASI)を修正したもので評価している。患者さんの返答が得られない場合は代理人に尋

ねているが、患者さんの返答率は 52%であった。代理人は患者さんが治療方針に対して意思決定できない場合の意思決定者とした。患者さん本人および代理人から情報を得られなかったときは(23%)、ADL および DASI は診断、年齢、合併疾患、Glasgow Coma Score, Acute Physiology Score and site を含む multivariate regression model に基いて算出した。

結果評価

結果は患者さんの生存期間、病院および ICU 入院期間、入院費、ケアの強さによって評価された。死亡日ないしは 180 日の時点での生存は全ての患者さんで確認された。入院費は退院時請求書より書く病院のコストとの比率より入院中のコストを割り出し、しかも 1993 年の値段に合わせた。一部の患者さんは ICU に入る前既に入院していたが、その際の費用もコストに含めてある。研究期間のコストを計算するために入院直後より研究に組み込まれた患者さんのデータが log total cost と平均 TISS の log と入院期間の関係をだすのに用いられました($R^2$ = 0.88)。この関係は全ての患者さんの研究期間におけるコストを算出するために用いている。ケアの強さは RHC のための TISS ポイントを除外した TISS で 1, 3, 7, 14, 25 日に決定している。

Data Quality

カルテ閲覧者とインタビュー・スーパーバイザーは中央にて訓練を受け、各施設をプロトコールがきちんと遂行されているかどうか年間 4 回訪問した。カルテの 10%をランダムに抽出し、第二のカルテ閲覧者によって独立して検討してもらい Acute Physiology Score では 89%の一致、TISS では 86%の一致率をみている。ランダムに抽出した７５のカルテにおける RHC 施行の有無に関しては 100%の一致率であった。インタビュー・スーパーバイザーが入力する前に間違いがないかもう一度検討している。結局研究対象としての適応がある患者さんの 81%にインタビューをすることができた。患者さんとコンタクトがとれないときの代理人のインタビュー回答率は 81%、医師は 87%であった。全てのデータは 2 回入力してもらい、National Coordinating Center にモデムとして送ってもらっている。Range check と internal consistency の check はデータ上で行なわれた。99%以上のカルテがプロトコール通りに記載されていた。

Propensity Score

解析前、critical care の 7 人の専門家に RHC を用いるかどうかを決定する際の因子をあげてもらった。これらの因子は 24 時間以内に RHC を行なわれたかどうかと合わせて多ロジスティック解析で算出された。独立した因子は年齢、性、人種、教育年数、年収、保険のタイプ、1 次的、2 次的病気のカテゴリー、入院時の 12 のカテゴリー、入院 1 週間前の ADL,DASI2、1 日目の蘇生術、癌の状態、SUPPORT モデルにおける 2 ヶ月以上生存する確率、APACHEHI score の急性生理部分、Glasgow Coma Score, 体重、体温、平均血圧、呼吸数、心拍数、$PaO_2/FIO_2$比、PaCO2, pH, WBC, Ht, Na, K, Cr, Bil, Alb, 尿量、合併症カテゴリーであった。多ロジスティック解析により RHC を行なう確率 (=propensity score: 0 - 1)として算出された。Propensity score は数多くある要素から RHC 適応をなるべく完全に予想するために計算されたが、その対象は 5735 人に限られており、必ずしも golden standard というわけではない。Propensity score が諸々の因子を表現するのに適当であるかどうかは RHC を行なった患者さんの propensity score を５つに分け、RHC を行なった人と行なわなかった人で個々の因子の相違を検討することにより決定した。

Case-Matching Procedure

RHC なしに治療された患者さんは病気のカテゴリーと propensity score に基き matching された。最初 2184 人の RHC を行なった患者さんの中からランダムに 1 人を選択し、RHC を行なわなかった 3551 人全ての中から同じ疾患カテゴリーで 0 から 1 の propensity score の中から 0.03 以内の違いまで許容してコントロール 1 人を選択した。この操作は全ての可能性のあるペアが同定されるまで行なわれた。最終的にそれぞれのペアの propensity score の最も近いものが選ばれました。最後に個々のペアにおける propensity score における相違を計算し、差がプラスのものと差がマイナスのものでうまく相殺し合うかどうか検討しました。

Sensitivity analysis

我々の得る所見の妥当性は RHC を行なったことに影響した全ての因子を調整する propensity score にかかっている。何故なら、今まで述べてきた方法を用いて、適当な全ての因子を propensity regression に含めて調整したが、重要な要素が propensity regression から抜けてしまっているリスクがある。この missing covariate を以下の

３つの sensitivity test で検討した。

この研究に関係のない 13 人の臨床医に RHC の適応を左右する思い付く因子を全て挙げてもらい、その中で特に重要な要素 10 を更に指定してもらった。この 10 の要素はすべて propensity score に含まれていた。我々は RHC 適応に最も重要な４つの因子を決定し、propensity regression からその 4 つのうちの１つを抜いたときの stability adjustment をみている。Rosenbaum と Rubin の提案した sensitivity test で missing covariate の影響を調べた。

解析

全ての解析は SAS ソフトウエアを用いて行なわれた。RHC を用いた群と用いなかった群の連続変数違いは Wilcoxon rank sum test で、カテゴリー的な違いは$\chi^2$ test で、mached pair の間の相違は signed rank test を用いて、30 日、60 日、180 日の生存は McNemar で評価された。生存曲線は log rank test で計算された。RHC と生存および病院滞在期間の関係は Cox proportional hazards model を用いて計算された。RHC と入院費用、平均 TISS の関係は linear regression model を用いて評価した。Subgroup analysis では Cox hazard model を用いて RHC と５つの施設における生存期間と予め特定した subgroup の関係を解析した。多重比較の補正は propensity によって層化された multiple covariate の間で RHC を行なったものと行なわなかったものとの間で解析した。

研究結果

study population について

9105 人の SUPPORT 患者さんのうち、5735 人は SUPPORT に組み込まれてから初日に ICU での治療を受けていまする。ICU 入院後最初の 24 時間以内に 2184 人が RHC を受けた。さらに 308 人が 72 時間以内に RHC を受けた。しかし 2184 人の RHC を行なった患者さんを行なわなかった 3551 人の患者さんと比較した。

Patient and disease characteristics

5735 人の特徴は Table 1 に示した。

Table 1. Characteristics of 5735 critically ill patients.

| 因子 | | No RHC (n=3551) | RHC (n = 2184) |
|---|---|---|---|
| 年齢 | - 49 | 884 (25) | 540 (25) |
| | 50 - 59 | 546 (14) | 371 (17) |
| | 60 - 69 | 812 (23) | 577 (26) |
| | 70 - 79 | 809 (23) | 529 (24) |
| | 80 - | 500 (14) | 167 (8) |
| 性別 | 男性 | 1914 (54) | 1707 (78) |
| | 女性 | 1637 (46) | 906 (41) |
| 人種 | 白人 | 2753 (78) | 1707 (78) |
| | 黒人 | 585 (17) | 355 (15) |
| | 他 | 213 (5) | 142 (7) |
| 疾患カテゴリー | 急性呼吸不全 | 1200 (34) | 589 (27) |
| | 多臓器不全 | 1245 (35) | 1235 (57) |
| | うっ血性心不全 | 247 (7) | 209 (10) |
| | 他 | 859 (24) | 151 (7) |
| 癌 | 無し | 2652 (75) | 1727 (79) |
| | 局所 | 638 (18) | 334 (15) |
| | 転移 | 261 (7) | 123 (6) |
| DNA status day 1, Yes | | 710 (20) | 296 (14) |
| APACHE III score (without Glasgow CS) | | 51[38, 50, 62] | 61[47, 60, 74] |
| 2 ヶ月以上生存する確率 | モデルよりの評価 | 0.61[0.49, 0.65, 0.76] | 0.56[0.45, 0.60, 0.72] |
| | 医師の評価 | 0.51[0.10, 0.50, 0.80] | 0.61[0.05, 0.50, 0.75] |
| 合併疾患の数 | | 1.7[1, 2, 3] | 1.6[1, 1, 2] |
| 2 週間前の ADL | | 1.6[0.5, 1.1, 2.4] | 1.5[0.5, 1.1, 2.2] |
| 2 週間前の DASI | | 20[16, 20, 23] | 21[17, 20, 23] |

| | | |
|---|---|---|
| 研究開始前の入院日数 | 4.6[0, 0, 5] | 6.4[0, 2, 8] |
| 体温 | 37.6[36.2, 38.1, 39.0] | 37.6[36.1, 38.1, 39.0] |
| 心拍数/分 | 112[76, 120, 140] | 119[105, 125, 145] |
| 平均血圧 | 85[53, 68, 119] | 68[47, 57, 73] |
| 呼吸数/分 | 29[20, 30, 39] | 27[12, 28, 37] |
| 白血球数 x $10^3$/・l | 15.2[8.2, 13.6, 19.4] | 16.2[8.6, 14.7, 21.2] |
| $PaO_2$ / $F_{IO2}$ | 240[149, 224, 333] | 192[110, 168, 267] |
| $PaCO_2$ | 40[32, 38, 44] | 37[30, 36, 40] |
| pH | 7.39[7.35, 7.40, 7.46] | 7.38[7.32, 7.40, 7.46] |
| Creatinin mg/dl | 1.9[0.9, 1.3, 2.0] | 2.5[1.2, 1.8, 3.0] |
| Albumin g/dl | 3.2[2.7, 3.5, 3.5] | 2.9[2.4, 3.5, 3.5] |
| Glasgow Coma Score | 11[7, 14, 15] | 10[4, 13, 15] |

患者数(%)、平均[$25^{th}$, $50^{th}$ (median), $75^{th}$ percentile],

DNR：do not resuscitate, APACHE：Acute Physiology and Chronic Health Evaluation, ADL：activities of daily living, DASI：Duke Activity Status Index, shade = $p < 0.01$

補正前の比較

RHC の患者さんは RHC を行なわなかった患者さんと比較して 30 日、60 日、180 日後の生存率は明らかに低く、研究観察開始日からの入院費も RHC 施行患者さんで有意に高い傾向にあった。RHC を行なった患者さんはより長く ICU ないし病院に入院する傾向にあった。

Table 2

| | | No RHC | RHC | p |
|---|---|---|---|---|
| 生存（%） | 30 日 | 2463 (69,4) | 1354 (62.0) | < 0.01 |
| | 60 日 | 2231 (62.8) | 1190 (54.5) | < 0.01 |
| | 180 日 | 1906 (53.7) | 1012 (46.3) | 0.01 |
| 保険使用率 | 費用(x$1,000) | 74.3 [18.4, 37.1, 81.5 | 131.9 [42.1, 81.7, 160.6] | < 0.01 |
| | 平均 TISS | 28 [21, 27, 35] | 35 [28, 35, 42] | < 0.01 |
| 入院日数 | ICU | 10.3 [3, 6, 11] | 15.5 [5, 9, 18] | < 0.01 |
| | 病院 | 20.5 [8, 13, 23] | 25.7 [9, 17, 32] | < 0.01 |

Propensity score

RHC を行なった患者さんと行なわなかった患者さんの propensity score はかなり重なっていた。RHC を行なった患者さんの propensity score の平均は 0.577 (95% CI 0.108 – 0.943)であり、RHC を行なわなかった患者さんの propensity score の平均は 0.253(95% CI 0.011 – 0.779)であった。RHC の重要な因子を調整するための propensity score の有効性については propensity を 5 等分してそれぞれの因子の相違を検討した。疾患重症度、平均血圧、心拍数、呼吸数、pH, PaO2/FIO2, PaCO2, 疾患分類を含む RHC の適応に関与する重要な因子は RHC を行なった患者さんの propensity score を 5 等分して検討したところ、RHC を行なった患者さんと行なわなかった患者さんで変わらなかった。（注＊このことは propensity score が同一であれば、RHC 施行、未施行にかかわ

らず予後因子は一致することを示している。しかしながら、あくまでも 5 等分なので、residual confounders を残す余地はあると考えるべきである。そのため Table 3 として propensity score を一致させて RHC, non—RHC を比較する際、重要な予後因子が均等に分布しているかどうかを確認した。）

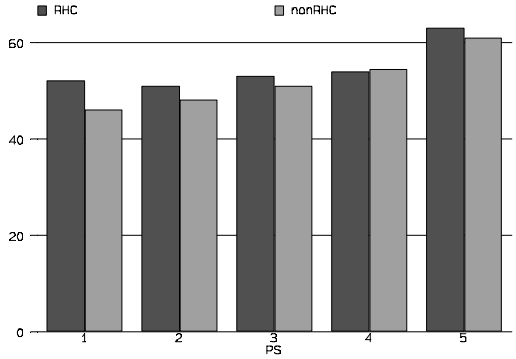



Figure 1-A: 横軸は propensity score を小さい方から５等分(Quintiles)し、RHC 施行群（黒）と非施行群で比較したものである。縦軸は Acute Physiology and chronic health evaluation (APACHE)III score 。

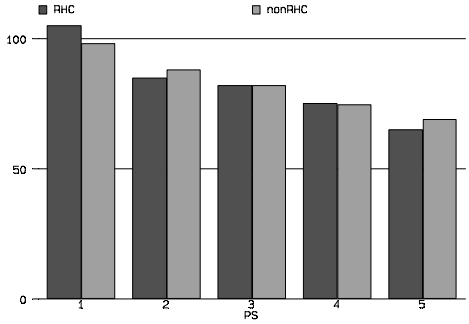



Figure 1-B: 上と同じ横軸条件で縦軸は平均血圧(mmHg)を示してる。

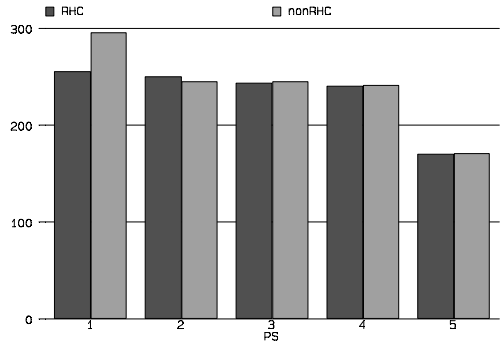



Figure 1-C: 上と同じ横軸条件で縦軸は $PaO_2/Fio_2$ (mmHg) を示している。

上の図をみて理解できるように、propensity score 分布は RHC 施行群と非施行群の間でほぼ均等であることが理解できる。すなわち「RHC は予後不良群に対して施行しているのだから、RHC 施行例で死亡率が高いのはやむを得ない」という説に疑問を投げかける形となった。例えば propensity score が高くなると血圧は低くなる傾向がある。若干 RHC 非施行群の方が高めの血圧であるが、ほとんど差は認められない。しかし上記は propensity score を５つに区切っただけなので、もっと細かく区切ったり連続変数として捉えたとき全く同じかどうかは疑問が残り、residual confounders の存在は否定できない。

Selection bias の調整：case matching (n=2016)

RHC と non-RHC の間で疾患カテゴリーと propensity score によって 1008 組を matching した（合計 2016 人)。

Table 3

| 因子 | No RHC (n=1008) | RHC (n = 1008) |
|---|---|---|
| Propensity Score | 0.51 [0.35, 0.50, 0.67] | 0.51 [0.36, 0.50, 0.67] |
| Acute Physiology Score (w/o Glasgow CS) | 57 [44, 58, 71] | 57 [43, 57, 70] |
| 年齢 | 60 [49, 63, 72] | 60 [49, 62, 73] |
| DNA status day 1, Yes | 710 (20) | 296 (14) |
| APACHE III score (without Glasgow CS) | 51[38, 50, 62] | 61[47, 60, 74] |
| 2 ヶ月以上生存する　モデルよりの評価確率 | 0.58[0.46, 0.62, 0.74] | 0.59[0.47, 0.62, 0.74] |
| 合併疾患の数 | 1.6[1, 1, 3] | 1.6[1, 1, 2] |
| 2 週間前の ADL | 1.6[0.5, 1.1, 2.4] | 1.5[0.5, 1.1, 2.2] |
| 2 週間前の DASI | 21[16, 20, 24] | 21[17, 20, 24] |
| 研究開始前の入院日数 | 6.8[0, 2, 8] | 6.5[0, 2, 8] |
| 体温 | 37.7[36.1, 38.3, 39.1] | 37.6[36.2, 38.2, 39.0] |
| 心拍数/分 | 111[105, 125, 145] | 111[103, 124, 145] |
| 平均血圧 | 73[49, 61, 108] | 71[49, 60, 81] |
| 呼吸数/分 | 28[19, 30, 38] | 28[14, 28, 38] |
| 白血球数 x $10^3$/・l | 15.3[8.2, 14.0, 20.0] | 15.0[7.4, 13.6, 20.0] |
| $PaO_2$ / $F_{IO2}$ | 210[127, 185.7, 296] | 211[120, 192, 305] |
| $PaCO_2$ | 37[31, 36, 41] | 38[31, 36, 40] |

| | | |
|---|---|---|
| pH | 7.39[7.34, 7.40, 7.46] | 7.39[7.34, 7.40, 7.46] |
| Creatinin mg/dl | 2.3[0.1, 1.6, 2.6] | 2.3[1.2, 1.7, 2.7] |
| Albumin g/dl | 3.0[2.5, 3.5, 3.5] | 3.0[2.6, 3.5, 3.5] |
| Glasgow Coma Score | 13[12, 15, 15] | 13[12, 15, 15] |

患者数(%)、平均[25th, 50th (median), 75th percentile],
DNR：do not resuscitate, APACHE：Acute Physiology and Chronic Health Evaluation, ADL：activities of daily living, DASI：Duke Activity Status Index, shade = p < 0.01

疾患分類は急性呼吸不全 46%、多臓器不全 34%、うっ血性心不全 11%、その他 10%であった。Table 3 に挙げた因子は RHC, non-RHC の間で統計学的に差を認めなかった。

マッチしたペア解析で、RHC を施行した患者さんの生存期間は調査した 30 日、60 日、180 日で明らかに低いことがわかった。

| 生存期間 | Non-RHC (n=1008) | RHC (n=1008) | OR (95% CI) | p |
|---|---|---|---|---|
| 30 日 | 677 (67.2) | 630 (62.5) | 1.24 (1.03 – 1.49) | 0.03 |
| 60 日 | 604 (59.9) | 550 (54.6) | 1.26 (1.05 – 1.52) | 0.01 |
| 180 日 | 522 (51.2) | 464 (46.0) | 1.27 (1.06 – 1.52) | 0.009 |
| 入院中死亡 | 629 (63.4) | 565 (56.1) | 1.39 (1.15 – 1.67) | 0.001 |

OR は 1.24 から 1.27 の範囲で RHC を行なった患者さんの方が早期に死亡していることが判明した。入院中死亡も RHC を行なった患者さんで 1.39 倍と高くなっている。30 日から 60 日までの生存曲線を比較したところ RHC 群が有意に劣っていた(p = 0.02)。RHC で治療を受けた患者さんの入院費および ICU 滞在期間は有意に高い値を示していた。

Table 5

| | | No RHC (1008) | RHC(1008) | p |
|---|---|---|---|---|
| 保険使用率 | 費用(x$1,000) | 35.7 [11.3, 20.6, | 49.3 [17.0, 30.5, | < 0.001 |

| | | | | |
|---|---|---|---|---|
| | | 39.2] | 56.6] | |
| | 平均 TISS | 30 | 34 | < 0.001 |
| | | [23, 29, 38] | [27, 34, 41] | |
| 入院日数 | ICU | 13.0 | 14.8 | < 0.001 |
| | | [4, 7, 14] | [5, 9, 17] | |
| | 病院 | 23.8 | 25.1 | 0.14 |
| | | [9, 15, 28] | [9, 16, 31] | |

治療 selection bias の multivariable analysis による調整(n=5735)

RHC を施行した患者さんについて propensity score だけでなく年齢、性、合併疾患、疾患カテゴリー、疾患重症度、Glasgow Coma Score, 入院 1 日目における 2 ヶ月後の予後判定、ADL, DASI についても調整したところ, RHC で治療を受けた後の患者さんの 30 日の時点での relative hazard of death は 1.21 (95% CI 1.09 – 1.25, p<0.001)であった。同様の調整を行なった後各疾患における relative hazard of death を算出したところ、急性呼吸不全: 1.30 (95% CI 1.05 – 1.61, p<0.001), 多臓器不全: 1.32 (95% CI 1.11 – 1.57, p<0.001)、 うっ血性心不全: 1.02 (95% CI 0.55 – 1.89, p=0.94)だった。その他の疾患では 1.06 (95% CI 0.80 – 1.41, p=0.67)であった。

Multivariable adjustment を行なってから RHC と保険との関連を調べてみました。RHC 施行群全てにおいて入院費(greater mean ± SE; $7,900 ±$3,900; p<0.001)、治療の強さ(TISS care) (7.0 ±0.3; p<0.001)、ICU 長期滞在(2.2 ±0.5; p<0.001)有意差を認めた。

Analysis by subgroup and by prognosis

RHC を行なった患者さんの 30 日の時点での adjusted relative hazard of death は老人、女性、白人、ショックあるいは敗血症のある患者さん、術後ケアを受けている患者さんで高い傾向にありました。逆に検討した全ての subgroup において RHC を行なう merit はなかった。

Sensitivity analysis

我々は不測の因子が治療 selection bias に及ぼす影響について調べてみた。最初に

13 人の臨床医に RHC の予後因子として考えられるものを１０挙げてもらったところ、全て propensity score の項目に入っていた。また propensity score の中で重要な項目(Pa02/FI02，平均血圧、心拍数、呼吸数)の１つを除外しても relative hazard of death は 0.01 しか変わらなかった。

研究結果に対する考察

我々は２つの方法で広範囲に及ぶ可能性のある諸々の交絡因子を調整した後、RHC が生存率等の結果にどのように影響したかについて検討した。最初に、RHC 施行群と非施行群で疾患カテゴリーと propensity score をマッチングさせた。予想された通り、このことによって RHC 群と non-RHC 群の間の予後因子の分布は同じになった。疾患と propensity score で control すると RHC 施行群の患者さんの特徴は matching マッチングした非施行群の患者さんのものと一致した。このように同じ特徴をもつ患者さんで RHC を施行された患者さんは明らかに高い死亡率、より強い治療、ICU により長く入院していた。

第二に、我々は 5735 人全ての中で、RHC 施行が及ぼす影響について multivariate analysis を施行した。Case-matched analysis でみられたように、propensity score で治療における propensity score を調整した後、RHC が死亡率を高めていることが明らかになりました。Subgroup analysis を行なっても RHC が予後を改善するという結果を得ることはできなかった。また上の解析と同様に重症の患者さんを RHC で扱うことにより、より強いケアを必要とし、ICU により長く入院し、入院費もよりかかるという結果を得た。

疾患 subgroup の中でも急性呼吸不全、多臓器不全に対して RHC を行なうと特に relative hazard of death が増加していた。

RHC が患者さんの予後を悪くする要因として、ライン由来の敗血症、心内膜炎、大きな静脈血栓等が考えられた。もしも医師のカテーテルの知識と技術が十分でないとしたら RHC の利点は見落とされるかもしれない。また、RHC は高い死亡率と高い医療費のかかる医療スタイルの１つの象徴ともとれる。これは急性心筋梗塞の際の RHC 使用とも似ている。確かに RHC 施行群では TISS score が高い傾向にあるが、これを除外しても RHC の患者さんに対するメリットは見出せなかった。RHC は多臓器不全の患者さんに対して心機能を助け酸素供給をよくするために行なわれるが、最初のランダム化臨床試験では

RHC 施行群で 67%死亡率が高いと報告されている。それでも医師の間では RHC が特殊な重症例に有効であると信じて行なわれていた。

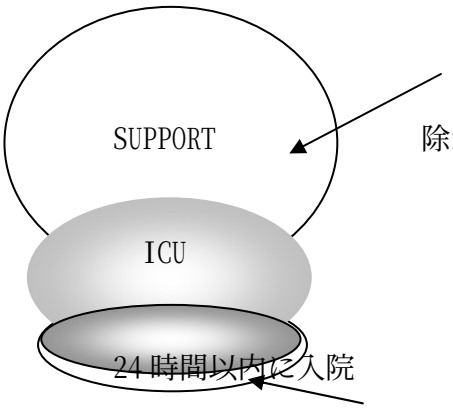
Limitation

本研究にはいくつかの重要な limitation がある。まず本研究は観察的研究でありランダム化臨床試験ではないということだ。我々は過去の報告よりも強力に選択バイアスを調整した。また不測の因子の影響も考慮したが、その可能性は低いと思われた。また我々は ICU に入院した患者さんで 24 時間以内に RHC を施行した人だけを対象としたが、24 時間以降に RHC を施行した人では何らかの利点を見出すことができたかもしれない。しかし RHC 施行群の 81%は 24 時間以内に行なわれていたのである。我々は本研究結果から RHC の施行は慎重であり滅多に使用するべきものではないと考えているが、本件急がランダム化臨床試験ではなく観察研究なので信用できないとするかもしれない。その中間の立場をとるのが最良かもしれない。

Study question に対する回答

１．この解析に含まれた患者さんは単に SUPPORT に含まれた患者さん全員の一部というだけだろうか？

SUPPORT に含まれた中でも最初の 24 時間に ICU に入院した患者さんのみを研究対象としている。患者さんは ICU 以外では RHC を受けていない。何故なら ICU に入院しない患者さんに対しては RHC の適応は基本的にないからである。また ICU にも入院せず、RHC も受けていない患者さんの propensity score は、RHC を受けたどの患者さんのそれより低いため matching を行なうことができない。この集団を含めた場合 propensity score の lowest quintile に含まれるが、これは residual confounding につながる可能性がある。

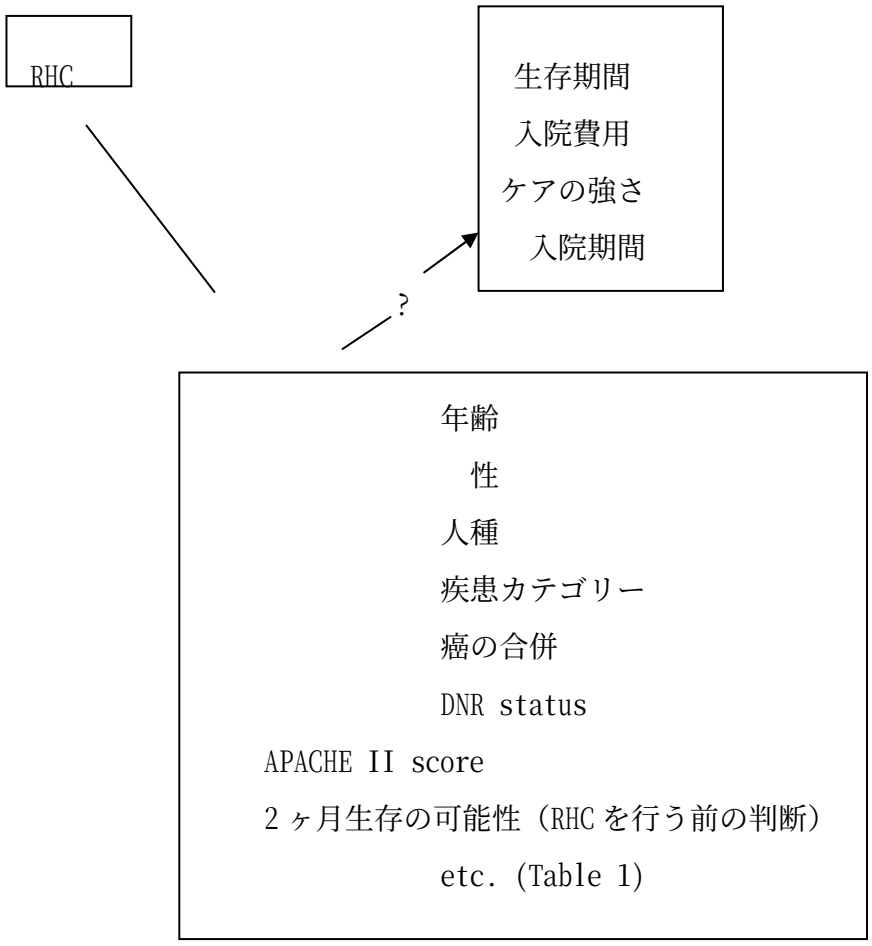



ICU に入院しなかった人は RHC の at risk ではないので除かれます。

RHC 施行患者はこの中に多くが含まれる(81%)

２．この研究における介入，結果，交絡について説明せよ。

介入：RHC，結果：生存期間、入院費用、ケアの強さ、ICU および病院の入院期間、交絡：病気の重症度、疾患分類、経済レベルなど諸々。Table 1 にあるリストは confounder となりえるが、結果との関連を示すデータがないため断定はできない。



３．7 人の専門化が RHC の可能性のある予測因子を同定しているが、本当は RHC 予測因子であるのに選択された予測因子から漏れてしまったものがあった場合、propensity score にどのような影響をもたらすか？

RHC を予測しえる全ての因子がみな propensity score に含まれている。研究者は何百もある全ての予後予測因子を一部に絞るよう専門家にアドバイスしているかもしれない。また漏れてしまった予後因子は交絡因子となりえる。

４．何人の患者さんがこのマッチングによる解析に組み入れられたか？何故減ってしまったのか？

各グループ 1008 人、合計 2016 人。これ以外の参加者については、propensity score0.03 以内のマッチングできるペアをみつけることができなかったから減ってしまった。

５．ICU 入院後 24 時間以上経ってから RHC を行なった患者さんは主な解析から外された。もし含めるとしたらどちらのグループに入れるべきか？晩期 RHC の二次的解析というのはどうか？

結果をみて eligibility を変えてはならない。よって 24 時間以降の RHC を含めなかったのはやむを得ない。しかし 24 時間以内に RHC を含めなかった患者さんは、適応があっても患者さんの具合が悪かった為に行なわれなかったかもしれない。そのような場合には 24 時間以降の RHC をも RHC 群に含めるべきだろう。24 - 72 時間の間に RHC を行なった患者さんを後で解析しているが、最初の 24 時間に RHC を行なわず後で行なった患者さんの人数は 3551 人中 308 人であった。著者らはこの 308 人を含めていた場合の解析を行なっていないため、結果に及ぼす影響は不明である。

６．RHC を行なった患者さんの propensity score の平均は 0.577 であり、95%信頼区間は 0.108 – 0.943 であり、行なわなかった患者さんのそれは 0.253 (95% 信頼区間 0.011 – 0.779) と記載されているが、正しい結果だと思うか？

そもそも propensity score は logistic regression model だから 0 から１の間である。しかも、解析した人数は非常に多いので 95% CI は狭いことが予想される。しかし、

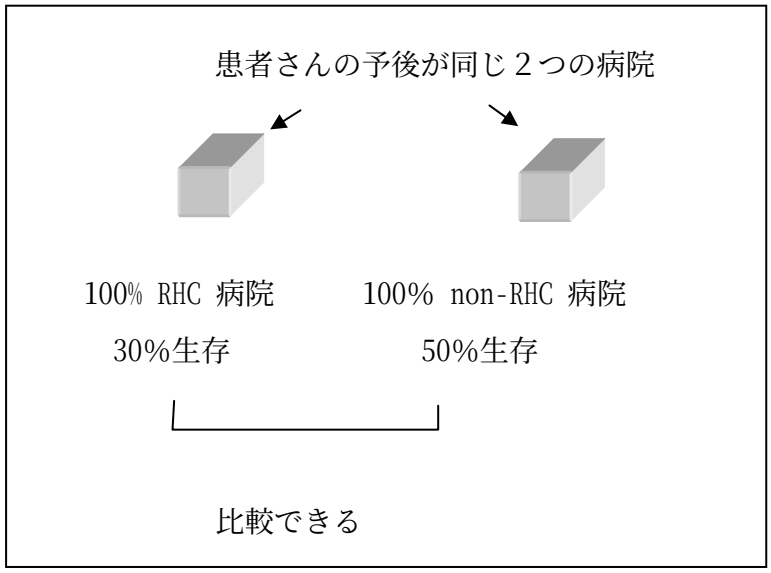
両者において 95%信頼区間は相当広く幅をもっている。しかも平均が 95%CI の中心には位置していない。信頼区間というよりは、最小値、最大値をとっていたかもしれない。

７．Propensity score に含めた予測因子のいくつかに対して、RHC 施行群と非施行群に分けて再度差がないかどうか検討しているが(Table 3)、なぜこのようなことをする必要があるのか？

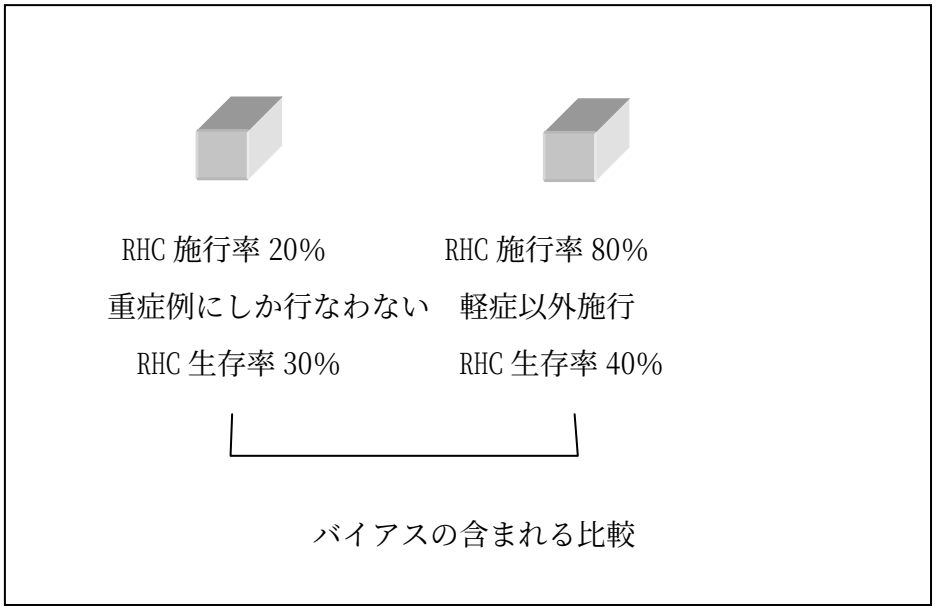
この表は、RHC, non RHC 間で因子の偏りがないかどうか確認するために作成された。この研究ではどの因子も 2 群間で偏りなく分布している。もしも偏りがあるとすれば、交絡因子となりえるため、更に多変量解析を用いた調整が必要となる。

９．　RHC を施行する病院は施行しない病院より死亡率が高いことになる。どのような状況下においてRHCと患者生存の関係をバイアスなく述べることができると思いうか？特に、RHC と基本的予後因子は比較された病院間でお互い独立している必要はあるだろうか？discussion において、著者らは RHC が単にケアの強い治療を好む医療のマーカーである可能性を示唆している。

RHC を行なう病院と行なわない病院で予後が変わらなければ問題ない。もちろん RHC を行なう病院で RHC をどの程度の頻度で行なうかに影響を受け、その割合が 0 or 100%であれば、重症度に関係なく RHC を行なうことになる。一方その中間であれば重症度とRHCの間に何らかの関係を生じ病院間でその関係が異なることが予想されるためバイアスを生じ帰無仮説の方向に近づく。この研究における[Hospital]は経済用語の「instrumental variables」に相当する。




個々の患者さんの予後に関係なく RHC の適応が決まっているため判断に迷うことがない。


RHC 施行率 20%
重症例にしか行なわない
RHC 生存率 30%
RHC 施行率 80%
軽症以外施行
RHC 生存率 40%
バイアスの含まれる比較

Propensity score ・ 多ロジスティック解析　どちらを選ぶべきか

図１．帝王切開は古くから行なわれてきた医療である。胎児仮死の徴候があれば、産科医は帝王切開に踏み切る。しかし、胎児仮死の徴候があれば出生時仮死のリスクも高くなる。そのため、一見帝王切開が仮死に結びつく錯覚におちいるが、さまざまな交絡因子が関与するため、そのような結論をくだすわけにはいかない。このような場合、仮死のあった児(ケース)と妊娠中毒症などの程度をマッチングさせて仮死の無かった児(コントロール）を選ぶマッチングという手法がある。一方、多変量解析により交絡因子の影響を補正する方法もある。このようにして純粋な帝王切開の仮死に及ぼす影響について知ることができる。

図２． Am J Epidemiol. 2003 Aug 1;158(3):280-7.
Am J Epidemiol に propensity score の適応例が示された。すなわち、結果発生数が変数の数の 7 倍以下である場合、propensity score を用いた方が的確な回答を得られるというのである。

図３．しかし、現実問題として仮死はそう多いものではない。今回の例では、4883 人の出産を調査し、176 人(3.6%) に仮死をみた。仮死に影響するのは妊娠中毒だけではなく、母親の年齢、分娩回数など多くの変数がある。具体的には結果発生（この例の場合仮死）数の数が変数Ｘ８より大きいとき、多変量解析で十分である。

図４．一方、結果を呼吸窮迫症候群に設定した場合、結果は僅か 46 人(0.9%) にしか認められなかった。結果に影響し得る変数を 13 選んだとすると、13 x 8 = 104 で 46 人より多くなる。このような場合、propensity score の方が精度が高くなる。実際、このように結果発生が少ない状況で多変量解析を行なうと、個々の因子の有意性がブレやすい。

図５．実際のお産のデータである。全部で 4883 例のデータがここにある。583 人が予定帝王切開を、430 人が緊急帝王切開を受けて生まれている。

図６．まず緊急、予定帝王切開が選ばれる因子を多ロジスティック解析で検討する。

図７．そして propensity score（PS）を図のような要領で算出する。

図８．帝王切開を行なわれた対象の propensity score は 0 から１まで広く分布している。一方、帝王切開を施行されなかった対象では、低い propensity score に集中している。

図９． Propensity score を 5 段階に分け、帝王切開の行なわれた頻度を調べてみた。Propensity score が大きくなるにつれて帝王切開が選択される割合が高くなっているのが理解できる。

図１０． Propensity score (PS) と帝王切開（allcaiser）をソートした EXCEL 表である。10 番目の対象は帝王切開を施行されているが、9 番目の対象は帝王切開を施行されていないので、これをマッチングさせる。ＰＳの値は両者で近似している。先の右心カテーテルにおいては 0.03 以内の範囲でＰＳのマッチングが行なわれた。

図１１.マッチングを行なったあと、帝王切開例 622 例、非帝王切開例 622 例となった。両者の propensity score の間に有意差はない。

図１２． Propensity score でマッチングする前は、妊娠中毒症の頻度が明らかに帝王切開例に多い（上)。しかし、マッチングしたあとは、均衡がとれ帝王切開施行例、非施行例の間で妊娠中毒症の併発に差を認めなかった。

図１３．マッチング前は、明らかに妊婦の年齢は妊娠中毒例で高い（上)。しかし、マッチングした後は近似し、両者の間で統計学的有意差を認めない。

図１４．帝王切開と仮死の発生について検討する。マッチング前（上）では、帝王切開により仮死が多くなると判定された。しかし、propensity score でマッチングしたあとは、仮死と帝王切開の間に有意な関係をみいだすことはできなくなっている。

図１５．今度は仮死の発生を通常の多ロジスティックモデルで解析してみた。

Propensity score を用いた場合同様、帝王切開は有意差を示していない。

図１６.今度は結果を仮死ではなく、ひり発生頻度の低い呼吸窮迫症候群(RDS)にした。マッチングを行なわないと帝王切開はＲＤＳの発生リスクを上げる。しかし。Propensity score でマッチングしたあとは、この有意差は失われる。

図１７．今度は多ロジスティック解析を行なった。前回の仮死の場合とは異なり、一番下の変数である帝王切開は有意性を示している。

以上のように、結果発生が変数 x７　より大きかった仮死の場合には、多ロジスティック解析を用いた場合にも、propensity score を用いた場合にも帝王切開が仮死に影響しないという点で一致していた。

しかし、呼吸窮迫症候群という少ない結果を選んだとき、その数は変数 x 7 を超え、Am Ｊ Epi の指摘した通りとなった。すなわち、「結果発生に対して選択する変数の数が極端に少ない場合、propensity score を用いるべきである」点である。