


厚生労働科学研究費補助金
政策科学総合研究事業
(政策科学推進研究事業)

The seal of the National Institute of Public Health is a large, circular emblem in the background. It features a central green stylized plant or flower design. The outer ring of the seal contains the text "NATIONAL INSTITUTE OF PUBLIC HEALTH" in a serif font, arranged in a circle.

治療中の者等に対する保健指導を行った場合の
医療費適正化の効果に関する調査研究

平成23年度 総括研究報告書

研究代表者 岡本悦司

(国立保健医療科学院)

平成24(2012)年5月

治療中の者等に対する保健指導を行った場合の 医療費適正化の効果に関する調査研究

研究代表者 岡本悦司 国立保健医療科学院 上席主任研究官
研究協力者 菊池清隆 全国健康保険協会 調査分析グループ

研究要旨

治療中者が特定保健指導を受けた場合の医療費への影響を全国健康保険協会のデータを用いて、プロペンシティスコア(傾向得点)マッチングとPDM法による傷病別医療費で評価した。糖尿病薬もしくは循環器薬を服用中でありながら特定保健指導を受けた群の累積医療費(調剤医療費含む)の伸びは、総医療費でもメタボ関連疾患医療費でも保健指導を受けた群は対照群よりも低く、特定保健指導は現に治療中の者に行うことで医療費伸び抑制効果があることが実証された。通常の特定保健指導の医療費への効果も含め、レセプトと健診データをリンケージし、適切なマッチングを行うことにより医療費への効果を評価する手法を確立した。

A. 研究目的

2008年に医療費適正化計画の一環として特定健康診査・保健指導が導入され、生活習慣病の発症予防により中長期的な医療費適正化が進められている。特定健康診査は40～74歳の全被保険者、被扶養者を対象とするが、特定保健指導は一次予防を目的とするので現に治療中の者はメタボ該当に階層化されても対象にはされない。しかしながら、医療費への効果を考えると健康者よりむしろ現に治療中者の重症化をくいとめた方が医療費制約効果は大きいのではないか、という考えもある。

本研究は、特定保健指導を治療中者にも実施したら医療費がどのように増減するかを評価することを目的とする。因果関係を評価する最も有力な方法は無作為割付比較試験(RCT)であるが、保険者の事業としてはそのような介入は技術的に困難であり、レセプトや健診データ等の既存データから推計するしかない。従来、既存データでは相関関係はわかっても因果関係まではわからない、が定説であったが過去数十年の間に、プロペンシティスコア(傾向得点)法という観察データでも一定の条件を満たせば因果関係を推定できる、ことが明らかになってきた。

本研究では、レセプトと健診データをリンケージし、プロペンシティマッチングという手法を適用して特定保健指導と医療費効果の因果関係を推論する。

特定保健指導は原則として治療中(服薬中)の者を対象としないことになっており、これに従うと治療中者には特定保健指導は実施されないため既存のレセプトや健診データからではこのような手法は理論上はとれないことを意味する。ところが実際には、治

療中か否かの判断はレセプト等によって確認するのではなく、特定健診の質問票において服薬の有無を受診者にたずね、受診者の回答でもって階層化している。受診者による回答は必ずしも正確ではなく、実際には治療を受けているにもかかわらず「服薬無し」と回答したため特定保健指導を実施された例が相当ある。本研究はそうした質問票の不正確さを逆に活用し、協会けんぽの大規模な被保険者集団をレセプトデータとリンクして治療中であるにもかかわらず特定保健指導を受けた被保険者を抽出し、それと傾向得点(プロペンシティブスコア)によりマッチさせた対照群(治療中でありかつ特定保健指導を受けなかった者)との間で保健指導前後の医療費の変化を比較することによって特定保健指導の医療費への効果を評価する。

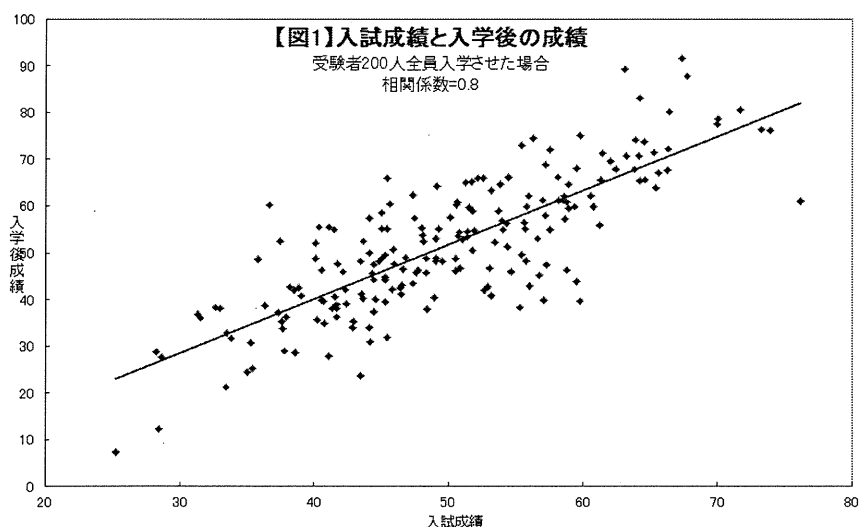
B. 研究方法

「レセプト情報・特定健診等情報データベース(通称ナショナルデータベース)」が2011年度より研究利用が可能となることから本研究もそれへの申出を行った。しかしながら行政機関個人情報保護法の制約によりプロペンシティブスコア(傾向得点)法のような分析は認められないとのことで却下された(資料は後掲)。幸い全国健康保険協会の理解と協力がえられたので協会のデータで同様の手法を試みた。協会は対象となる40~74歳人口の約4割をカバーしており、十分な標本数を有する。

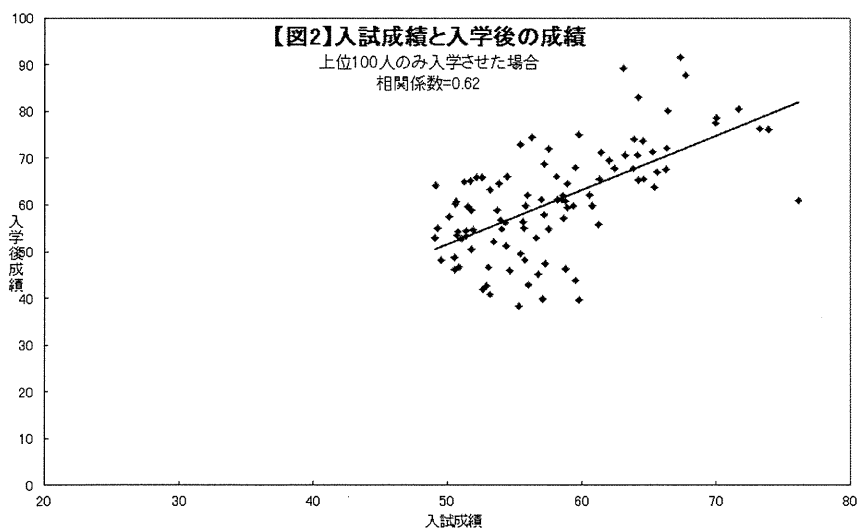
●プロペンシティブマッチングの原理

プロペンシティブスコア(傾向得点)法を数学的に説明することは複雑なので、わかりやすく説明する。観察研究においては相関関係はわかっても因果関係はわからない、が従来の定説であった。その最大の理由は観察研究では、全てのケースを観察できない、からである。

例として「入試成績と入学後の成績」の相関を知りたいとする。下のように入試成績と入学後の成績には高い相関があったとしても通常はこのパターンを観察することはできない。なぜなら、入試成績下位の者は入学できないので下の「真」の図を観察することは不可能だからである。



実際には、200人の受験者のうち上位100人のみ入学させ、入学者について入試成績と入学後の相関をみると下図のように弱い相関しか観察されない。本当の相関は受験者全員を入学させなければわからない、とされる。



●Rosenbaum & Rubinによる発見

しかし1983年にRosenbaum & Rubinがプロペンシティブスコア(傾向得点)法を考案したことから、観察研究でも因果関係を推計できることが明らかとなり、次第に定着しつつある。

たとえとして有名大の入試合格は所得にどう影響するか?を考える。

国民の所得を調査して、有名大卒1500万円、その他国民500万円と知ることができる。だからといって学歴の効果は1000万円とはいえない。なぜなら難関の入試を突破した人は頭もいいし努力もするだろうからフツウの国民とは異なる人種である。だから単純な引き算(1500-500=1000)を学歴の効果とはいえない。

ここで、A君、B君二人の受験生を考える。両君は共に入試前に模擬試験を受けて共

に合格可能性 60%と判定されていた。しかし A 君は合格、B 君は不合格……。その後、両君を追跡したところ A 君の卒後年収 1500 万円、B 君は 1000 万円であった……。このような場合、学歴の効果は 500 万円といえる。なぜなら、入試合格という介入は AB 両君の間で、一種無作為に割りつけられた、と考えられるからである。

この模擬試験における合格可能性がプロペンシティブスコア(傾向得点)である。重要な点は、A 君は英語に強く、B 君は数学が強い、というふうに二人は決して同一ではないが、しかし合格可能性が同じ者は入試合格という割付けにおいて同一と見なすことができ、それゆえ、合格者と不合格者の差異は、合格という介入による効果と判断できる。模擬試験における英語や数学という科目は健診項目のような共変量に相当する。

● マッチング

マッチングとは、できるだけ「似通った」者同士を組み合わせることである。性、年齢だけなら 24 歳男性を二人探しだすことはさほど困難ではない。しかしながら、変数(共変量)が多くなるとピタリ一致を探すことは困難になる。性、年齢に加えて、所得、学歴、体重、身長、喫煙本数そして HbA1c に血圧までピタリ一致する二人を発見することはきわめて困難である。研究デザインにおける階層化とは、共変量が一定範囲内の者に対象を限定してしまうことであるが、これでは得られた結果が普遍化できない。共分散分析による補正は、目的変数と共変量との関係をモデル化する必要がある。

先の入試と模擬試験の例では、英語、数学、国語とあらゆる科目の得点がピタリ一致する二人を探しだすのは困難だが、合格可能性という一つの数値に近い者を探しだすことは比較的容易である。介入群か対照群かの割り付けのプロペンシティブスコア(傾向得点)を対象者について算出し、プロペンシティブスコア(傾向得点)というひとつの数値が同じか近い者同士を組み合わせれば、たとえ組み合わせられた者同士は異なっても、組み合わせられた 2 群全体の共変量(所得、学歴、体重、身長、喫煙本数、HbA1c、血圧等)の分布はほぼ一致する。再び入試と模擬試験を例に出すと、合格可能性が同じなのに合格した者と不合格だった者とをくみあわせれば、組み合わせられた一人一人の各科目の成績は異なっても、組み合わせられた集団全体の各科目の得点分布はよく一致する。だとすると組み合わせられた合格者と不合格者の集団は同一集団とみてさしつかえない。

実験的研究においては無作為割付と呼ばれるが、観察研究においてはバランスよく(共変量の分布が一致するように)マッチングされた場合「強く無視できる割付け(strongly ignorable assignment)」条件を満たす、と表現する。介入群と対照群いずれかへの割付はロジスティック回帰に含めた共変量のみで決まり他の要因は「ほとんど無視できる」、というニュアンスである。これは、プロペンシティブスコア(傾向得点)マッチングを実施する上で重要な条件である。

再度、入試の例を出すと、可否は試験の成績だけで決まることが前提である。もし模擬試験の科目に入試科目が含まれていなかったら正しい合格可能性を判定することは

できない。また合否が試験成績だけでなくたとえば親の財力が影響するといった裏口入学や、A0入試等でボランティアやスポーツ歴も合否で考慮される、となると模擬試験の成績だけでは当然ながら合格可能性を正しく判定することはできない。

●プロペンシティマッチングを適用できる条件

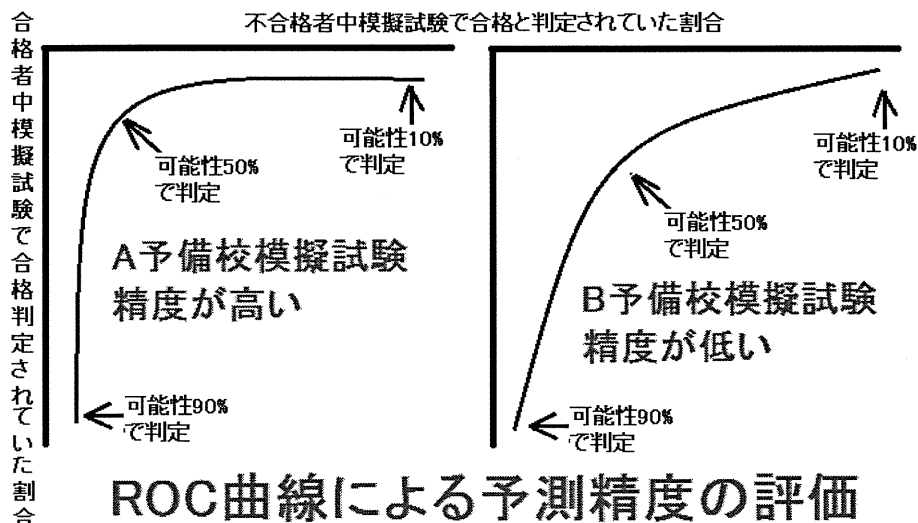
プロペンシティマッチングを適用する上で、確認しなければならない条件が3つある。それらは1)判別精度、2)適合度そして3)バランスチェックである。以下、これらについて詳説する。

1)判別精度(discrimination)

また模擬試験によっても合否判定の精度は異なる。判定の精度は一般にROC曲線を用いて評価される。入試は受けるか受けないかの二者択一であり、受けるかどうかの判定を模擬試験の合格可能性で判定する時、判定基準を甘くしてたとえば可能性10%以上の者全員を合格判定者として受験させると合格者に占める模擬試験での合格判定者の割合は高まるが、逆におびただしい不合格者を出すことになる。逆に判定を厳しくして可能性90%以上の者しか受験させないと受験者の大半は合格するが、合格者中に占めるその予備校出身者は少なくなり、また受験すれば合格したはずの多くの者のチャンスを奪うことになる。

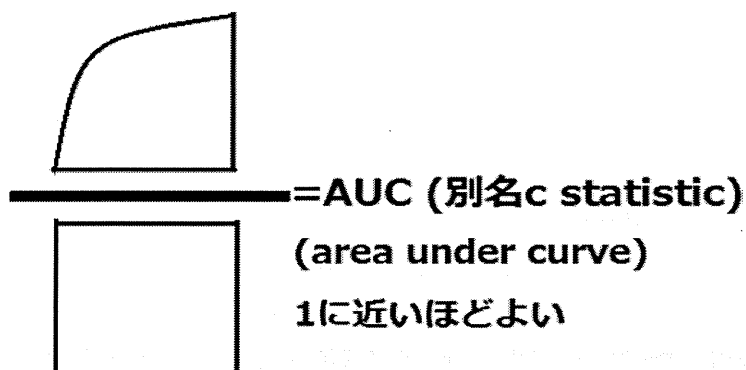
このように判定基準を甘い→厳しいに変化させ、合格者中の合格判定されていた者の割合をタテ軸、不合格者中合格と判定されていた者の割合をヨコ軸に描いたものがROC(Receiver Operating Characteristic)曲線である。受信者動作特性と訳されるのは、第二次大戦中、原始的なレーダーの画像から飛来する影が敵機か鳥かを肉眼で判断するレーダー手の精度は決定的であり、レーダー手の精度を評価する目的で開発されたためである。今日では特に各種検査の精度評価に用いられる。

ROC曲線が左上の角に近いほど精度は高いといえる。

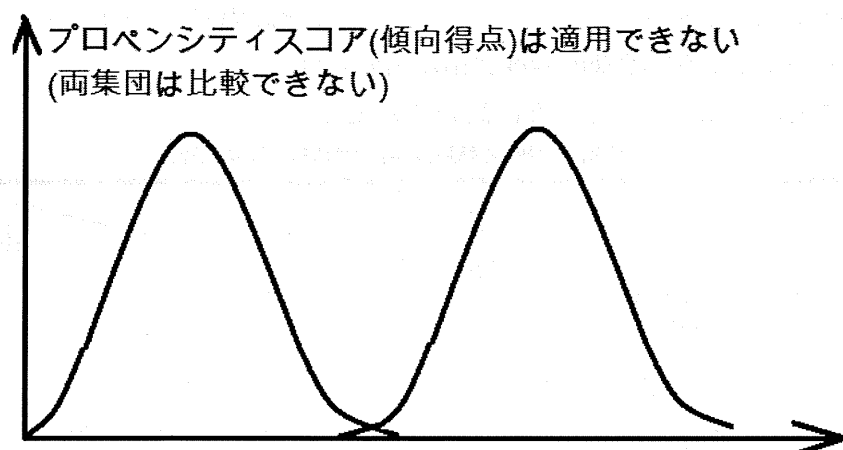


A予備校の模試の方がB予備校の模試よりも合否を正確に判定できることは一見して

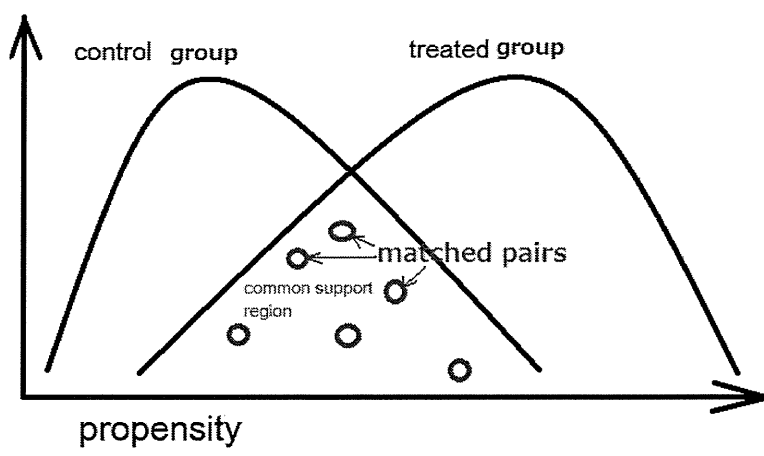
明らかであるが、それを定量的に把握するためにカーブ下面積 (Area under curve, AUC, c statistic と呼ばれることもある) が用いられる。 $1 \geq AUC \geq 0.5$ の値をとり、0.5 では全く判定に役立たず (サイコロを転がすのと同じ程度の正確さ)、1 に近いと百発百中ということになる。



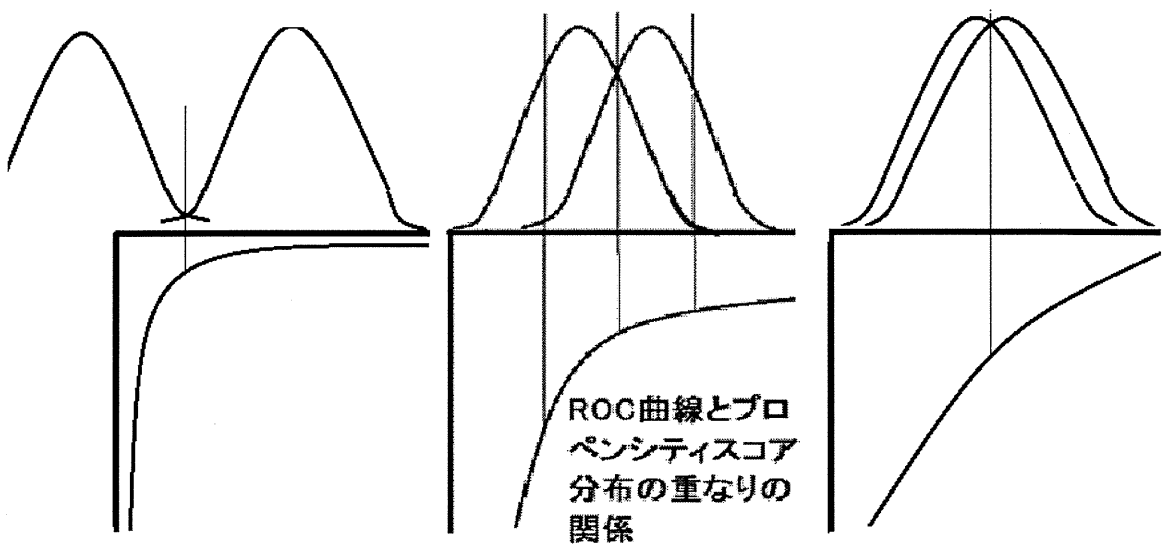
AUC が 1 に近いほど精度は高いのだが、ではプロペンシティブスコア (傾向得点) マッチングにおいても 1 に近いほどよいかというところでもない。入試と模擬試験の例でもわかるように、合格可能性と同じと判定されながら、合格と不合格とがある程度混在していないとプロペンシティブスコア (傾向得点) を用いることはできない。模擬試験の精度は高い方がよいが、逆にあまりよすぎて百発百中に合否を予測できたらプロペンシティブスコア (傾向得点) は用いることができないというジレンマがある。AUC が 1 に近いと下図のように両群間に重なりはなくなるのでプロペンシティブマッチングはできないことになる (もっとも 1 に近い AUC など実際にはめったになく問題になることはまずない。実際には AUC が 0.7 もあればオンの字である)。



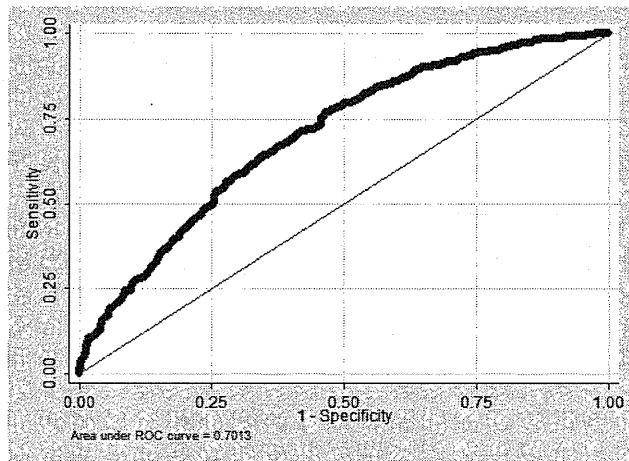
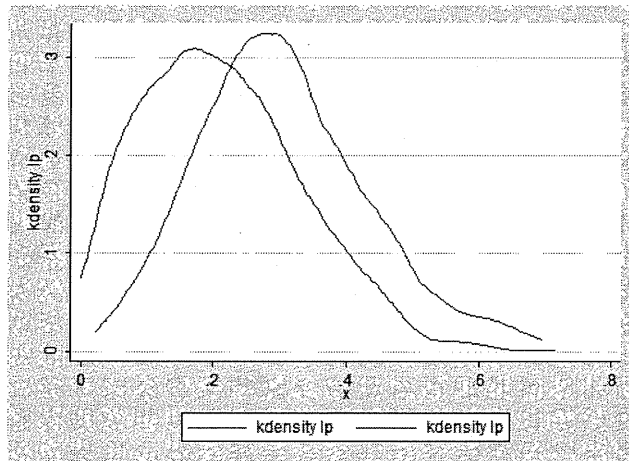
以下のように両群にはある程度の「重なり (common support region と呼ばれる)」が必要である。STATA 等のソフトでは common support を指定することにより重なり部分だけから介入群と対照群を抽出することができる。ただし、下図から明らかなように使用できる標本数は少なくなる。



AUCは1に近いほど判別精度は高いことを意味するが、プロペンシテイスコアマッチングの目的では、あまり1に近過ぎると common support region を確保できない、という問題も生じる。もし上記のような common support region にこだわるとしたらいっそう下図右のように完全にオーバーラップしていた方がよい、という見方もできる。実際にはAUCが1に近いことはめったになく、かといって0.5では両群は完全にオーバーラップしていることになり、AUCについては0.6から0.7くらいが限界ではないかと思われる。



某健康保険組合の実際の STATA による分析結果は以下の通り。AUC0.7 と理想的な数値である。



2) 適合度 (goodness of fit)

採択されたロジスティックモデルがどれだけ実際の数値に適合しているかを評価する。評価には通常 Hosmer-Lemeshow 検定が用いられる。Hosmer-Lemeshow 検定とは、プロペンシティブスコアを小→大に 10 群にわけ、それぞれの群ごとの期待数(モデルによる予測数)と実際に観察された観測値とを比較し、カイ二乗検定(自由度 8)を行う。STATA では下のように表示される。一番下の Prob>chi2=の値が 1 に近いと適合度は良いと判定される。

Probit model for hokenshido, goodness-of-fit test

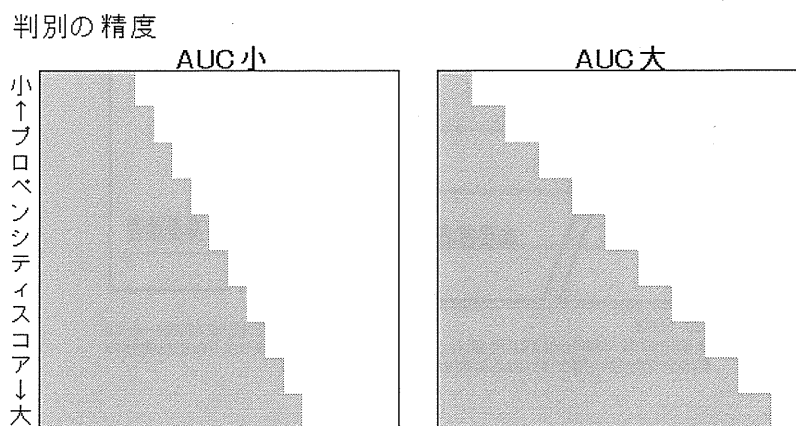
{Table collapsed on quantiles of estimated probabilities}

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0706	5	7.0	166	164.0	171
2	0.1177	16	16.0	154	154.0	170
3	0.1562	20	23.2	150	146.8	170
4	0.1899	32	29.4	138	140.6	170
5	0.2222	39	35.0	131	135.0	170
6	0.2592	40	41.0	130	129.0	170
7	0.2922	51	46.9	119	123.1	170
8	0.3412	55	53.7	115	116.3	170
9	0.4072	59	62.0	111	107.0	170
10	0.7174	80	81.9	90	88.1	170

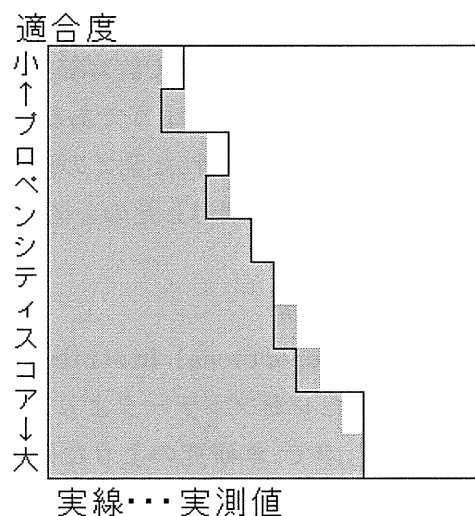
number of observations = 1701
 number of groups = 10
 Hosmer-Lemeshow chi2(8) = 3.00
 Prob > chi2 = 0.9243

精度(AUC)と適合度の関係を整理する。後述のように、今回の分析では適合度(HL 値)は1に近かったが、AUCは0.6未満という結果であった。理解のため両者の関係を整理する。

まずAUCは判別の精度であり、保健指導を受けるか受けないかを明確に判別できるかの指標である。HL表を使うと下のようにプロペンシティブスコア小と大との間の差が大きいことを示す。



Hosmer-Lemeshow 検定は、プロペンシティブスコアで予測された予測値と実際の結果がどれだけ近いかの評価である。いくらAUCが1に近くても、Hosmer-Lemeshow 検定の結果、有意な差がある(=実際と適合していない)というのではダメである。



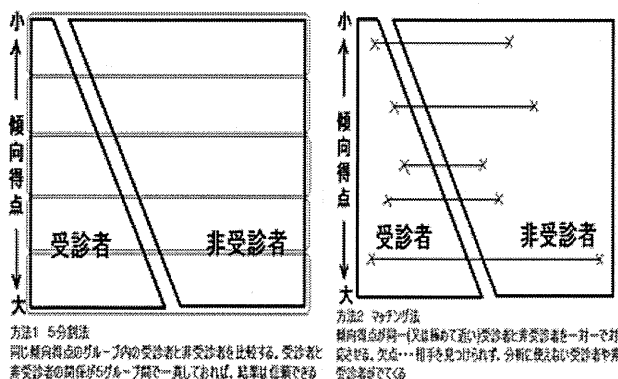
それゆえ、AUCとHosmer-Lemeshow 検定の両方をバランスよくみたまモデルを選択することが重要といえる。

3) マッチングとバランスチェック

マッチングには2つの方法がある。ひとつは階層化法で対象者全員をプロペンシティブスコア(傾向得点)の大から小へ5群に分割し観察するもので、もうひとつは一対一マッチングである。前者は標本の全てを使用できるので標本数は確保できるが、必ずしもプ

ロペンシテスコア(傾向得点)が近くない者も含めてしまうという問題がある。その点、
 一対一マッチングは正確であるが、マッチできない標本もでて、標本数を犠牲にしなけ
 ればならない、という欠点もある。

同一のプロペンシテスコア(傾向得点)の者同士で比較する



介入群と対照群に割り付けたら最後の仕上げは、バランスチェック(同一であることの確認)である。評価は標準化差すなわち両群の平均値の差を標準偏差で割ったもので行うべきとされる。T検定は用いてはならない。なぜなら、T検定では標本数が大きくなると小さな違いでも有意差がでるからである。バランスチェックは標本数とは無関係でなければならない、その意味でも標準化差を用いる。

標準化差はむしろゼロが理想だが、がどれだけまで許容範囲かというコンセンサスはない。一応全て0.25以内に納まっておけばOKのようである。また標準化差は、ロジスティック回帰に投入した変数だけでなく入手された全ての変数について確認すべきとされる。両群が同一であればあるほど無作為割付したのと同じ結果になる。

● 比例配分(PDM)法による傷病別医療費推計

傷病別医療費の推計には比例配分(Proportional Distribution Method, PDM)法を用いた。PDM法は多数の傷病が記載されたレセプトデータより傷病ごとに客観的な重みを測定して日数、点数を比例配分するもので、本研究のような傷病別の医療費の変化を客観的に評価する目的では、恣意的に主傷病を一つ選択しそのレセプトの医療費を全てその傷病のための医療費と見なす従来の主傷病法よりも優れている。とりわけ本研究のように3年近い期間にわたって、同一人の医療費を継続的に観察する場合には、従来の主傷病法では同一人のレセプトがある月のレセプトはたとえば糖尿病が主傷病、翌月は高脂血症というふうに異なる分類が行なわれたりすると医療費推計に大きな狂いが生じるだけに、PDMの価値は大きい。

PDMは従来よりフリーウェアとして公開提供されてきたが、電子レセプトに対応した

構造になっていなかったため、今回新たに電子レセプト形式に対応した Version4 を開発し使用した。その概要は末尾にマニュアルを添付した。また本研究で開発された Version4 はサンプルデータとともにネット上で公開されており、誰でもダウンロードして使用することができる。くわしくは下記サイト参照されたい。

<http://resept.com>

●データ抽出

全国健康保険協会が保有するレセプトデータならびに特定健康診査・保健指導データより以下のように抽出した。

対象

- H21年4月～H23年10月まで協会けんぽの継続加入者であること
- H22年3月末において40～74歳（特定健診の条件）
- 積極的又は動機付け支援に該当する者
- 健診の3月前から前月までの間に糖尿病又は循環器の治療（入院、入院外、調剤レセプトで判定）を受けている者（調剤および医科に糖尿病薬又は循環器用薬を含むもの）
- H21年4月～H22年3月において人工透析患者でない者
→人工透析は人工腎臓の他に、「薬効分類番号34の薬」、「特記事項02、03、16+慢性腎不全」で判定した。

介入群

- 母集団のうちH21年4月以降に保健指導を受けた者（保健指導を受けた時期は問わない）

対照群

- 母集団のうち保健指導を受けなかった者

STATAで対照群を抽出するために必要となる項目

- 性別、年齢、本人家族別、問診票の22項目、身長、BMI、収縮期血圧、中性脂肪、HDL、LDL、GOT、 γ GTP、空腹時血糖、尿糖、尿蛋白、所見
- 健診受診者の年齢はH21年度末年齢とし、母集団の年齢から平均と標準偏差を計算し、標準正規化した
- 検査値（尿糖、尿蛋白以外）の欠損は母集団における性別平均値とする→抽出後Excelでプロビット変換（ $\text{NORMSINV}(\text{PERCENTRANK}(x, \text{範囲}, \text{有効桁数}))$ ）を使用する。PERCENTRANKの有効桁数は10桁程度とし、NORMSINVは $\text{PERCENTRANK}(x, \text{範囲}, \text{有効桁数})=0, 1$ のときエラーとなるため、0には0.00001を加算し、1には0.00001を減算して対応した）
- 問診票（尿糖、尿蛋白を含む）は2値変換（欠損値は0）。3つ以上の選択肢のある質問については以下のように2値変換した。
 - ・食べる速度は「普通、遅い」→0、「速い」→1
 - ・飲酒は「ほとんど飲まない」→0、「毎日、ときどき」→1

- ・飲酒量は「1合未満」→0、「1～2合、2～3合、3合以上」→1
- ・生活習慣改善は「改善するつもりはない」→0、その他はすべて1
- ・尿糖、尿酸は「－」→0、その他はすべて1

【医療費集計】

- 総医療費、入院外医療費（PDM用）、調剤薬剤料が必要
- 標本それぞれの医療費集計期間は初回面接月の当月～H23年10月診療分まで（標本により期間が異なる）とし、経過月ごとの医療費が必要
- 対照群の初回面接月は対応する介入群と同じとする
- 総医療費は入院外と調剤の合計（全レセプト）
- 入院外（PDM用）は電子レセとOCRレセ（傷病は3個まで）を対象とし、レセプトごとの日数、点数（対応する調剤レセプトの点数を加算）、傷病コード（121）が必要
- 調剤薬剤料は薬効分類コード3桁が396（糖尿病薬）、214（降圧剤）、218（高脂血症）のものを集計

（倫理面への配慮）

本研究で用いたレセプトならびに特定健康診査・保健指導データは全国健康保険協会がその業務のために保有するものであり、統計法ならびにナショナルデータベースのような行政機関個人情報保護法の適用を受けていない。本研究は研究代表者が全国健康保険協会との間で覚書を締結し、研究代表者が企画立案し統計ソフト等を提供し、協会の調査分析グループがその業務の一環として実施した。よって集計された結果表等のみ研究代表者に提供され、個人情報にあたるデータは提供されていない。

C 研究結果

●選択された共変量

特定健診を受診した者のうち糖尿病薬又は循環器薬を投与されていた者は26121人いた。その中3521人が特定保健指導を受けていた。STATAのpsmatch2というプロペンシディマッピング用のソフトを用い、最善の判別精度ならびに適合度が得られる共変量を選択した結果、以下のような共変量が選択され同数の対照群が選択された。

Probit regression

Number of obs = 26121
 LR chi2(27) = 394.39
 Prob > chi2 = 0.0000
 Pseudo R2 = 0.0191

Log likelihood = -10131.115

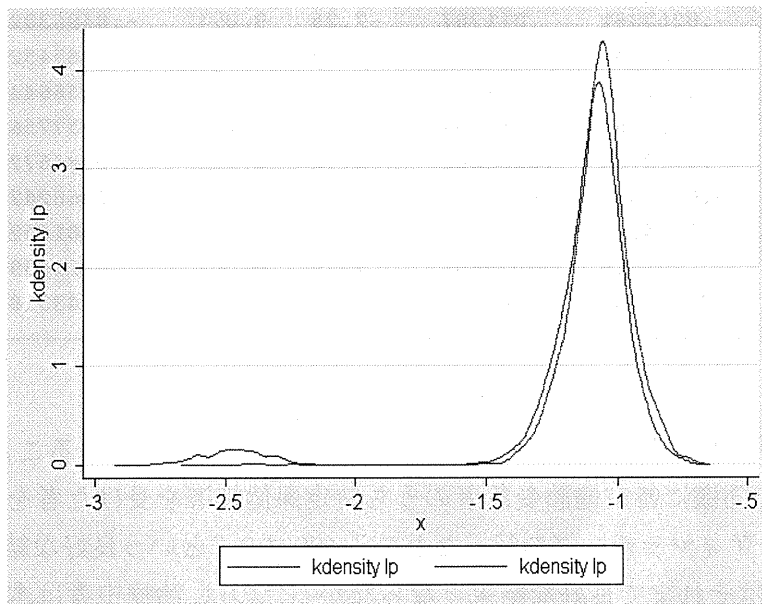
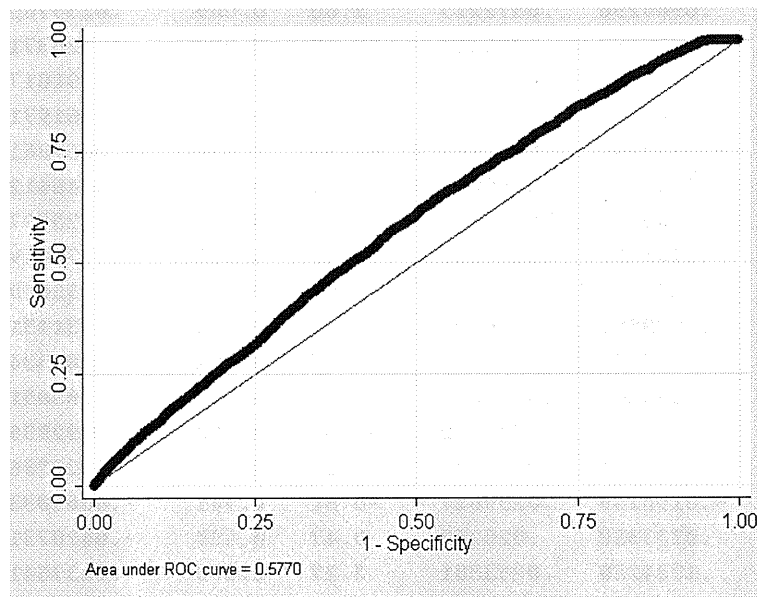
HOKENSHIDO	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
AGE	-.0415117	.0109801	-3.78	0.000	-.0630324	-.0199911
SEX	.0966296	.0372215	2.60	0.009	.0236769	.1695824
HONNINKAZOKU	1.382433	.1306463	10.58	0.000	1.126371	1.638495
PHXAMI	.1396282	.0359835	3.88	0.000	.0691017	.2101546
PHXDIALYSIS	-.2062586	.1814893	-1.14	0.256	-.561971	.1494539
SMOKING	-.0432965	.0221481	-1.95	0.051	-.0867061	.000113
WTCHANGE20	-.0237184	.0259815	-0.91	0.361	-.0746412	.0272044
EXERCISE30M	-.065587	.0341704	-1.92	0.055	-.1325597	.0013856
ACTIVITY	-.0181987	.0300446	-0.61	0.545	-.0770849	.0406876
WALKINGSPEED	-.0438151	.0262217	-1.67	0.095	-.0952088	.0075785
WTCHANGE1Y	-.0089824	.0276335	-0.33	0.745	-.0631431	.0451783
EATING1	.0212826	.0253148	0.84	0.401	-.0283334	.0708987
EATING2	-.0560401	.0268092	-2.09	0.037	-.1085851	-.0034951
EATING3	-.0917532	.0344602	-2.66	0.008	-.1592939	-.0242125
EATINGHABIT	-.0857321	.0341825	-2.51	0.012	-.1527286	-.0187355
INSHURYO	-.0158149	.0259027	-0.61	0.541	-.0665833	.0349535
KAIZEN	.0133332	.028269	0.47	0.637	-.0420731	.0687395
WILL	.1334258	.0253381	5.27	0.000	.0837641	.1830875
HEIGHT	-.039629	.0133882	-2.96	0.003	-.0658694	-.0133885
FFA	-.0265361	.011787	-2.25	0.024	-.0496382	-.0034339
HDL	-.0159643	.0115211	-1.39	0.166	-.0385452	.0066167
LDL	.0141887	.0101293	1.40	0.161	-.0056644	.0340419
GTP	-.0130661	.0115867	-1.13	0.259	-.0357756	.0096434
FBS	-.0178755	.0105062	-1.70	0.089	-.0384672	.0027162
USUGAR	-.1090445	.0498885	-2.19	0.029	-.2068242	-.0112647
UPROTEIN	.0065206	.0279014	0.23	0.815	-.0481652	.0612063
SHIDOKUBUN	.1507049	.0558893	2.70	0.007	.0411639	.260246
_cons	-2.667703	.1406222	-18.97	0.000	-2.943318	-2.392089

●判別精度

AUCは0.577にとどまった。いくつか試みたがAUCが0.6を超えることはなかった。健康保険組合で同様に特定健診受診者のうち特定保健指導を受けた者を特定健診データでプロペンシティマッチングを行ったところAUCは0.7という良好な結果がえられているのに、今回はそれまで良好な結果とならなかったのは、治療中者は本来は特定保健指導の対象ではなく、質問票の記載誤りといった偶発的に(何かの間違いで)特定保健指導を受けてしまったため、正確に予測することが困難であるためかもしれない(通常の特設保健指導は特定健診データと質問票に基づきシステムティックに対象者が選定さ

れる)。

もっとも、前述のように、AUC が 1 すなわち、介入群と対照群がプロペンシティブスコア (傾向得点) で完全に判別される (=両者に重なりがない) とプロペンシティブスコア (傾向得点) を実施することはできない。しかし AUC が 0.5 では特定保健指導を受ける予測として全く役立たない。今回の 0.6 という AUC は不満足なものではあるが、プロペンシティブマッチングの妥当性を確保できるギリギリであったといえる (事実、AUC 0.52 で国際誌に掲載された例もある)。



● 適合度

Hosmer-Lemsho 適合度検定を行った。26121 人を 10 グループに特定保健指導を受ける確率の小→大に分け、それぞれのグループにおける実際に保健指導受けた数 (観察数)

とモデルで予測された予想数がどれだけ一致するかをカイ二乗検定で評価するもので、もし完全に一致したら1になる。下のように1に近い良好な結果が得られた。

前述のように高い判別精度はプロペンシティマッチングでは必須ではないが(AUC が1に近いとプロペンシティマッチングは困難だから)、高い適合度は必須であり、プロペンシティマッチングによりバイアスが十分に調整されたことを示唆する。

Probit model for HOKENSHIDO, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

小 ↑ 保 健 指 導 を 受 け る 確 率 ↓ 大	観察数		予想数		Total	
	Group	Prob	Obs_1	Exp_1		Obs_0
1	0.1014	144	145.2	2469	2467.8	2613
2	0.1164	291	286.7	2321	2325.3	2612
3	0.1261	312	317.6	2300	2294.4	2612
4	0.1332	344	339.1	2268	2272.9	2612
5	0.1392	347	355.9	2265	2256.1	2612
6	0.1449	380	370.9	2232	2241.1	2612
7	0.1509	396	386.1	2216	2225.9	2612
8	0.1587	412	403.9	2200	2208.1	2612
9	0.1708	418	428.7	2194	2183.3	2612
10	0.2608	477	487.0	2135	2125.0	2612

number of observations = 26121
 number of groups = 10
 Hosmer-Lemeshow chi2(8) = 1.86
 Prob > chi2 = 0.9850

ピッタリ一致したら1 ←

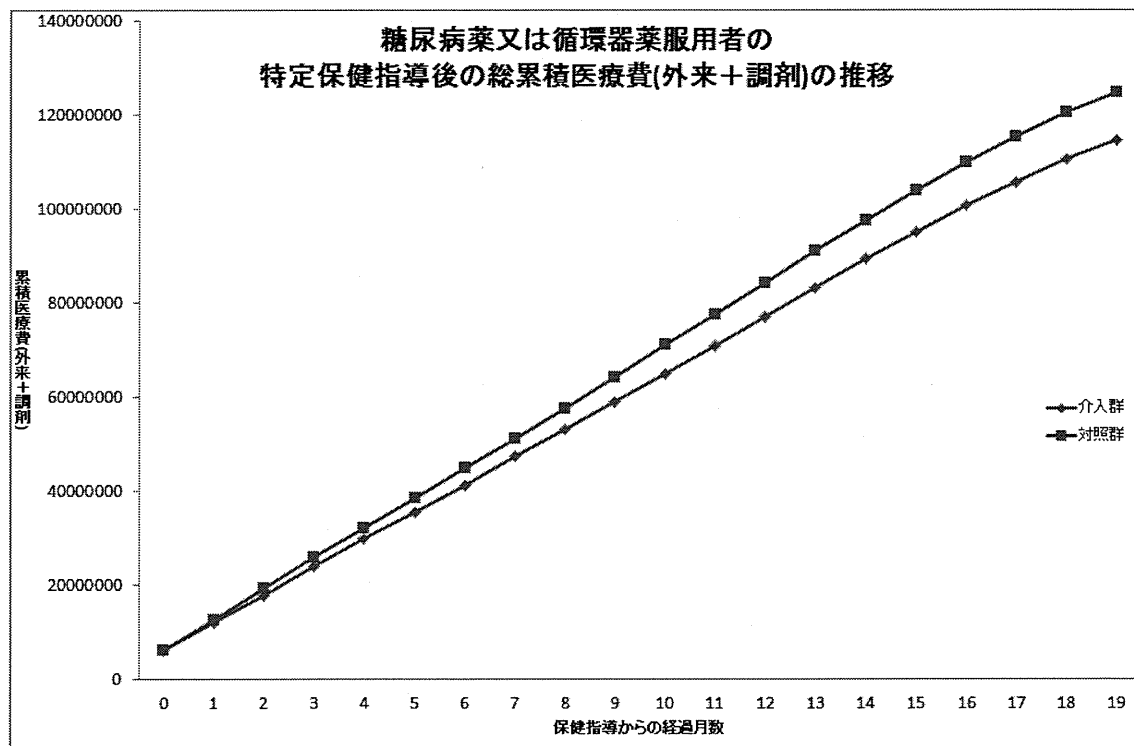
● バランスチェック

本人家族のひとつを除いて標準化差は小さく、両群は同一と考えることができる。

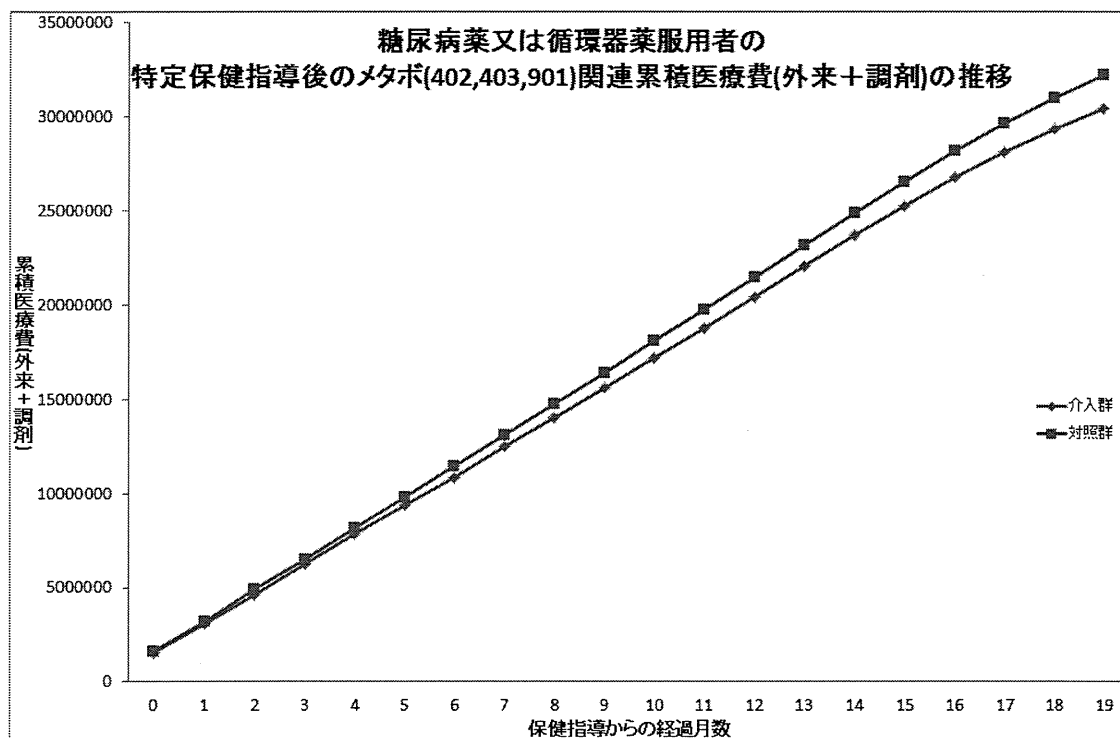
	Mean in treated	Mean in Untreated	Standardised diff.
AGE	-0.06	0.01	0.069
SEX	0.04	0.01	-0.069
HOKENKAZOKU	1.00	0.95	-0.971
PEXAMI	0.10	0.08	-0.057
PHDIALYSIS	0.00	0.00	0.025
SMOKING	0.32	0.33	0.021
WTCHANGE20	0.44	0.45	0.022
EXERCISE30M	0.11	0.13	0.061
ACTIVITY	0.16	0.18	0.047
WALKINGSPEED	0.24	0.26	0.039
WTCHANGE1Y	0.19	0.20	0.019
EATING1	0.27	0.27	-0.007
EATING2	0.21	0.23	0.035
EATING3	0.10	0.11	0.059
EATINGHABIT	0.10	0.11	0.051
INSHURYO	0.27	0.27	0.011
KAIZEN	0.51	0.50	-0.003
WILL	0.35	0.32	-0.056
HEIGHT	0.02	-0.01	-0.032
FFA	-0.04	-0.00	0.033
HDL	-0.06	-0.03	0.031
LDL	0.00	-0.02	-0.020
GTP	-0.03	-0.02	0.005
FES	-0.08	-0.03	0.049
USUGAR	0.04	0.05	0.064
UPROTEIN	0.15	0.15	0.008
SHIDOKUBON	0.97	0.96	-0.073

● 累積医療費の比較

両群の総累積医療費(外来+調剤)の推移は以下の通り、介入群の医療費は一貫して対照群より低く、その差は拡大している。



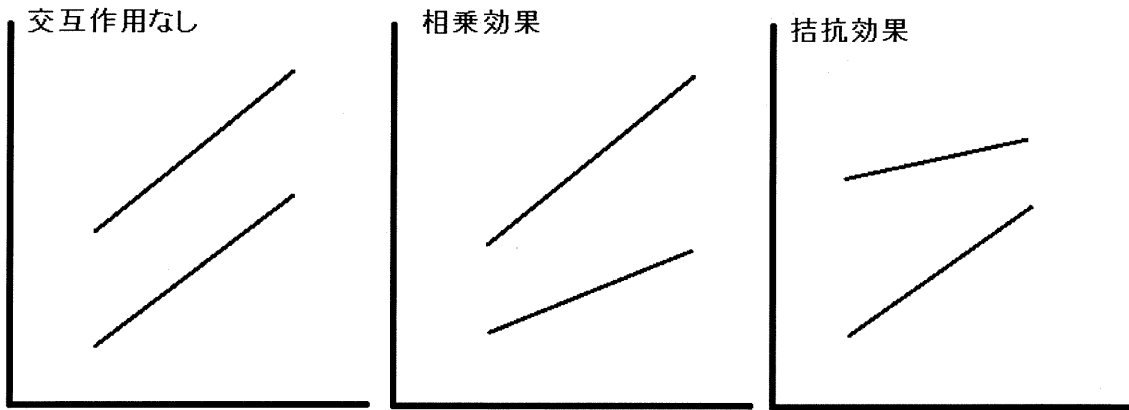
PDM法によりメタボ関連の3傷病分類の医療費のみを抽出した。メタボ関連3傷病分類(糖尿病 402, 高脂血症 403, 高血圧 901)は外来+調剤医療費の約26%を占めていた。総医療費と同様にその累積額は介入群の方が対照群よりも低く、月が経過するごとに差は拡大していた。



● 有意性の検定

上のグラフからわかるように、総医療費でもメタボ医療費でも特定保健指導を受けた介入群の累積医療費の伸びは対照群を下回っていた。また月が経過するにつれて両群の差は拡大するように見える。この所見が標本数の小ささによる偶然のものでないか有意性の検定を試みた。

両群の回帰直線の「傾き」が同一なのかどうかは交互作用の有無の検定になる。すなわち指導後の経過月数と特定保健指導の有無という二つの因子がからみあって相乗効果あるいは拮抗効果(それぞれ正・負の交互作用とよぶ)があるかどうかの検定である。



経過月数と保健指導の有無の交互作用を含むモデルを下のような統計ソフト R のプログラムで行った。

```
model<-lm(TENSU~KEIKA*HOKENSHIDO)
を実行し,
summary.aov(model)で結果を抽出した。
```

総医療費の場合

共分散分析結果(総医療費)

	自由度	偏差平方和	偏差平方和の平均	F値	p値	Pr(>F)
経過月数	1	495900000000000000	495900000000000000	51272	2.00E-16	***
保健指導の有無	1	304400000000000000	304400000000000000	314.7	2.00E-16	***
経過月数と保健指導	1	104200000000000000	104200000000000000	107.7	2.28E-12	***
その他	36	348200000000000000	96720000000000			

メタボ医療費の場合

共分散分析結果(メタボ関連医療費)

	自由度	偏差平方和	偏差平方和の平均	F値	p値	Pr(>F)
経過月数	1	339800000000000000	339800000000000000	72738.1	<2e-16	***
保健指導の有無	1	719000000000000000	719000000000000000	153.9	0.000000	***
経過月数と保健指導	1	245200000000000000	245200000000000000	52.5	0.000000	***
その他	36	168200000000000000	467100000000			

上記のように、総医療費でもメタボ関連医療費でも経過月数と保健指導の有無は正の交互作用(=相乗効果)があり、すなわち保健指導による医療費抑制効果は一時的なものではなく、時間がたつにつれてますます増強してくることが示された。

D 考察

中長期的な医療費適正化を目標に開始された特定健康診査・保健指導であるが、特定保健指導による生活習慣の改善によって医療費にいつ頃どのような影響がでてくるかをどのようにして評価するかは、これまで明確な手法は確立されてはいなかった。各保険者がみずからの保健指導の評価を正確に行えるように明確な手法の確立がまたれていたところであった。

本研究において、保険者が自ら保有するレセプトデータと健診データをリンケージすることにより科学的に妥当な評価を行う手法を提示することができた。具体的には、

- 1) 特定健診受診者の中で保健指導を受けた者とそうでない者をプロペンシティブスコア（傾向得点）法によって同数マッチングを行う。今回は治療中者を対象としたので、特定健診受診者でありかつ、健診の3か月以内に糖尿病または循環器薬がレセプトに出現している者にしぼった。
- 2) プロペンシティブマッチングは、特定保健指導を受ける受けないを目的変数としてロジスティック回帰を行い、特定保健指導を受ける確率の等しい者同士同数組み合わせる。なお特定健診データには欠損値が多く STATA のマッチングプログラムではエラーがでるが、これは性別の平均値を挿入することで回避する。また質問票のうち回答が3枝以上のものは2値(0又は1)に変換し、数値はprobit変換により正規化する(Excel関数ではNORMSINV(PERCENTRANK(x, 範囲))で代用できる。
- 3) プロペンシティブマッチングはSTATAという統計ソフトのPSmatch2というコマンドで実行できる。特定健診データと質問票のデータを説明変数(共変量)として用いるが、変数を取捨選択して繰り返して行い、1)判別精度(ROC曲線下の面積すなわちAUCの値)、2)適合度(Hosmer-Lemeshow)そして3)両群バランスを観察する。PSmatch2というソフトはこれらを同時に行うことが可能であるが3つとも最適なものを得ることは困難である。たとえば1)の判別精度が高い(AUCが1に近い)ことは両群の重なりが少ないことを意味し、もし重なりがないとプロペンシティブマッチングそのものを行うことができなくなる。よって、2)適合度が高いものをまず探し、次いでAUCも適度に高いものを選択する。最終的に3)をpblchkで確認する。
- 4) 選択された対照群と介入群について、レセプトを抽出し、介入後の経過月数ごとの累積医療費を算出する。ここで対照群として一対一で選択された者については、相手となった保健指導受診者と同時に同じ保健指導を受けたものとみなして経過月数を算出する。医療費は総額とともに、レセプトに記載された全傷病を用いたPDM法によって傷病別の累積医療費も算出して比較する。
- 5) 累積医療費の差が統計的に有意(偶然によるものではない)なことを検定する。そのためには累積医療費の2群×経過月数間のバラツキ(偏差平方和)を群と経過月数ならびに群×経過月数の3部分に分解し、群×経過月数のバラツキが有意に大きいかで検定する。有意に大きければ、両群の直線の「傾き」は異なると判断できる。