

生命保険数理における医療データ活用 - 因果探索とその応用

PwC Japan 監査法人

板倉 兼介君

PwC Japan 監査法人

堤 康平君



滝瀬 はい、おはようございます。それでは、定刻になりましたので、セッションC-1の「生命保険数理における医療データ活用－因果探索とその応用」を開始いたします。本セッションの司会を担当させていただきます、朝日生命の滝瀬と申します。よろしくお願いいたします。

本セッションでは、PwC Japan 有限責任監査法人の板倉さんと堤さんのお二人より、生命保険において、医療データを利活用する際に因果探索を活用する手法をお話しいたします。板倉さんは、日本アクチュアリー会正会員、CERAであり、PwCの保険アシュアランス部門ディレクターとして、保険会社・公共クライアントの医療データの利活用支援に注力しています。堤さんは、板倉さんのチームにおいて、アクチュアリー、データサイエンティストとして活躍されています。

それでは、板倉さん、堤さん、よろしくお願いします。

板倉 はい。滝瀬さん、ありがとうございます。

皆様、おはようございます。PwCの板倉です。対面で皆さんにお会いできることをとても楽しみにしておりました。とともに、オンラインでも昨日、「視聴しますよ」とたくさんお話をいただいて、非常に緊張しておりますが、どうぞよろしくお願いいたします。

本日は、最近、アクチュアリー会の方でもよく耳にされるといいますし、アクチュアリー以外の保険数理の文脈でよく使われている因果探索という手法をご紹介したうえで、ベイジアンネットワークという特定の手法をご紹介して、保険引受リスク評価で活用するというを試みたいと思います。比較的、非伝統的な技術を使うということになりますので、そういったものを取り込むときに、実務的にどのような結果と向き合っていくか、どのようなところで苦労するのかをお伝えできればよいかと思っています。

前半・後半ございますが、ご質問については最後にお受けしたいと思いますので、よろしくお願いします。

アジェンダ

- 1 課題設定
- 2 因果探索と
ベジアンネットワーク
- 3 ベジアンネットワークの
DeSCデータへの適用例
- 4 本日のまとめ
- 5 Appendix

PwC

2

では、本日は、このような流れでご説明します。最初に、1で課題設定、2で、先ほど申し上げた因果探索とベジアンネットワーク、3で、今回の因果探索の結果、手法を加味して、分析の設計や、あと結果の考察を、実践的にDeSCヘルスケア社さんのデータに使ってみたというものをご紹介します。最後に、今回の取り組みを、実務面等も踏まえて振り返りたいと思います。お時間の関係で、非常にテクニカルなところはAppendixに記載しております。ご興味のある方は、どうぞご覧ください。

アジェンダ

- 1 課題設定
- 2 因果探索と
ベジアンネットワーク
- 3 ベジアンネットワークの
DeSCデータへの適用例
- 4 本日のまとめ
- 5 Appendix

PwC

3

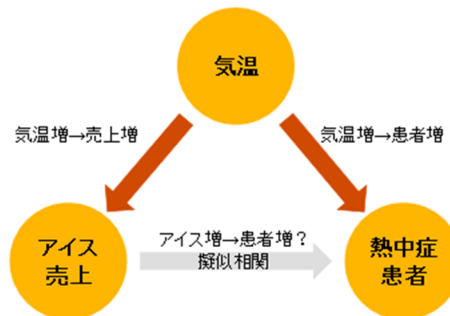
それでは、最初の課題設定というところから参ります。

因果関係とは？

因果関係とは、原因が結果に与える影響を指す。介入とは、原因を変化させて結果に与える操作とし、介入効果とはその影響として定義する

例:

「アイスの売上と熱中症患者の数」



PwC

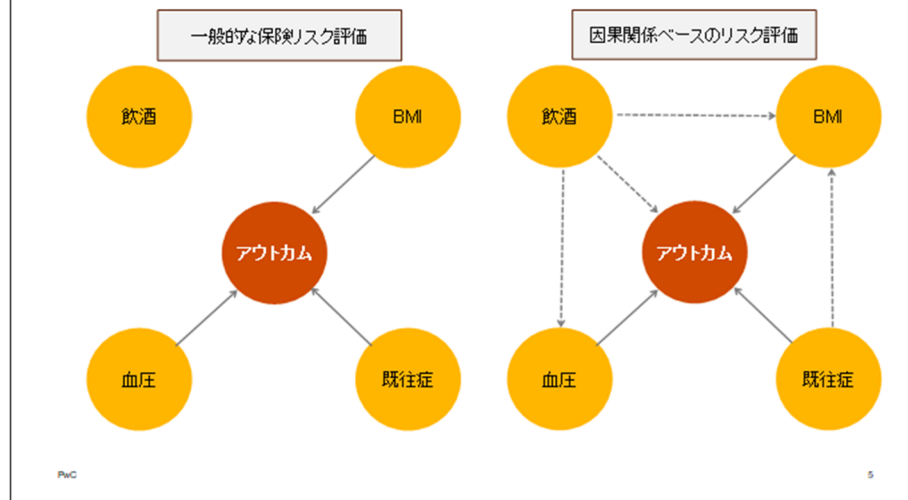
4

まず、「因果関係」や「介入」という言葉の定義を確認したいと思います。「因果関係」とは、原因が結果に影響を与える関係といったものを指していまして、「介入」で原因を変化させて、結果に与える操作というものを意味するとします。例えば、飲酒習慣とBMIや、お酒を飲む回数が増えたら体重が増えるなど、そういった割と当たり前の因果のようなものが想像できるかと思うのですが、少し注意しなければいけない例として、今、絵でお示ししているような、アイスの売上と熱中症患者の数といった関係を考えてみましょうか。

この場合、アイスの売上を増やすことで、熱中症患者の数が増えるといったことがあるのでしょうかというように考えるのですが、実は間に、別々に気温というファクターがあって、このファクターがそれぞれの変数に因果関係を持っているというようなことが推察されます。こういった見かけ上の因果関係が生じる現象のことを、「疑似相関」といった言葉で呼ばれます。「気温によって交絡が生じている」というように呼びます。この介入という、アイスの売上を操作するようなことを考える場合には、交絡因子と呼ばれるもののからの影響が混入してしまうといったことにご注意いただければと思います。

保険リスク評価における課題

引受リスクに関するリスク因子評価を単純加算して保守的に評価するのが一般的だが、実際はリスク因子間には因果関係が存在した上でアウトカムへ影響を与えている



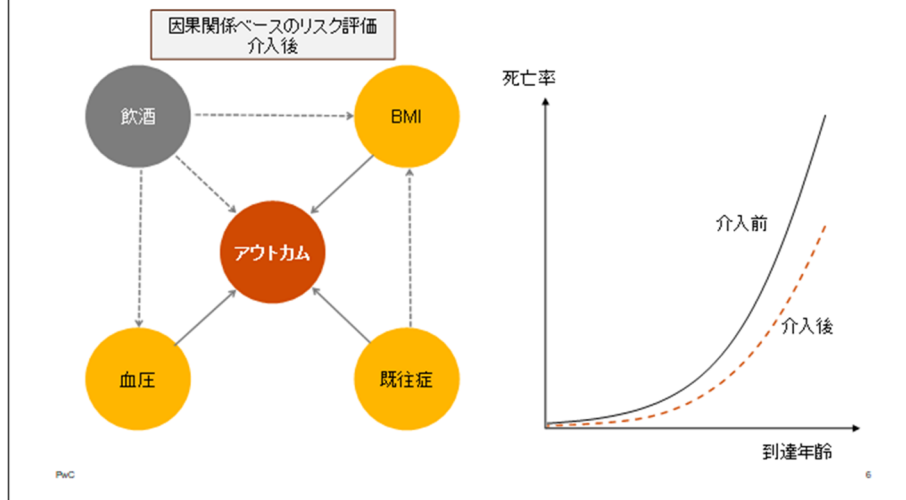
では、今日の課題設定というところです。まず、左側をイメージしていただきたいのですが、こちららは一般的な保険引受リスク評価、従来型のリスク評価を表しています。シチュエーションを明確にするために、新契約の引受査定を、BMIや血圧、既往症といった情報を基に行うことを想定していただければと思います。真ん中の「アウトカム」というものは変数なのですが、こちらは死亡など、支払事由に相当するような変数と読み替えていただければいいかと思います。

こちらは、通常は個別項目ごとに申込者の体況を評価した点を加算していきます。例えばBMIでいえば、高いから、少し太っているからプラス50と。血圧でいえば、血圧が少し高いのでプラス50。合計プラス100で、標準体よりも死亡率を2倍にして保険料を提示しますといったようなことが、実務的に行われています。また、それもよく知られていることなのですが、こちらのような評価を加点方式といっていますけれども、保守的にしすぎてしまっていて、リスクの実態に見合っていないような保守性を持っています。

一方で、右側の方なのですが、こちらはリスクファクター間に、実はこういった因果関係が存在しているといったことを表しています。実際にこういったリスクファクターを評価したうえで、アウトカムにどれだけ影響があるのかを評価したり、では、このリスクファクターを変えていけば、アウトカムの置きやすさにどう影響するのかを評価していきたいということが、今日のテーマになります。

因果関係導入の利点

因果関係を考慮したリスク評価を導入する事で、観察データをもとにリスク評価が可能になり、最良推定前提を押し下げる介入効果を定量的に測定する事ができる



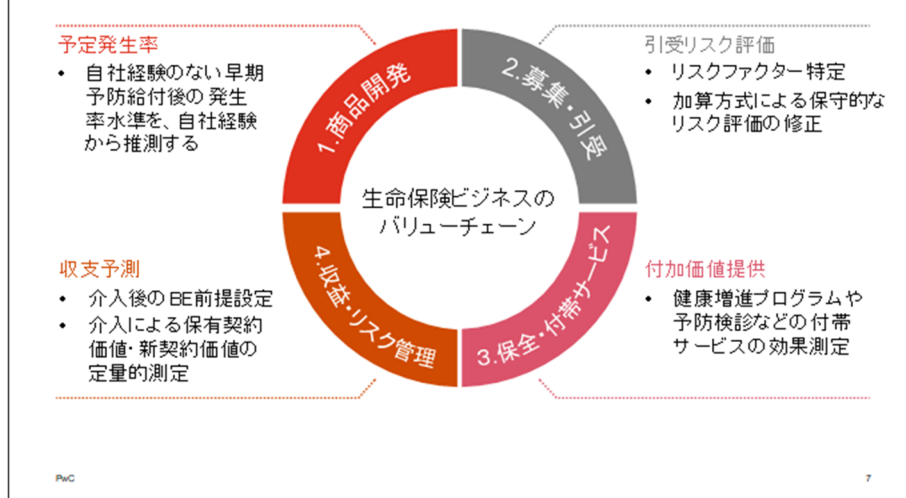
このページでは、では、そのメリットは何かというところを明確にご説明します。

まず、少し申し上げたのですが、飲酒量や頻度など、先ほどここは点でつながっていなかったのですが、そこを考慮に入れた保険リスク評価ができるといったことが一つ。

もう1点は、リスクファクターには、先ほど申し上げた、こういったドット・ラインがあると、因果関係があるというようになっています。例えば、飲酒習慣を改善したらBMIが改善して、血圧も改善して、最終的にアウトカムが発生する確率も下がるなど、そういったことが起きるのだといったことを、今、こちらの右側の絵で表示していますが、それはアクチュアリー言葉に直すと、「ベスト・エスティメイト前提が、介入によって下がるのですよ」といった表現ができると思っています。こういった介入の効果を定量的に表現していきたいといったことが、今日、できることとしてお伝えしたいことです。

活用シーンの例

観察データから因果探索・因果推論を行う事で、生命保険ビジネスのバリューチェーンのあらゆる場面での活用を期待できる



次のこちらのページで、この考え方は、では、生命保険ビジネスのバリューチェーンで、どのようなとこ

ろで活用できるのかといったところをご説明しているのですが、「あらゆるところです」と一言で言えます。特に通常のバリエーションの目線と異なるところは、ベスト・エスティメイトの前提は、スタティックで動かせないものだと見るのが一般的だとは思いますが、今日のお話で、それに介入していくことで、実はダイナミックに動かしていくような前提だと見なせるのではないかということを、ご提案したいといった趣旨になっています。今日の、先ほどちらりとご説明したユース・ケース、適用例では、この右側の二つをご紹介します。

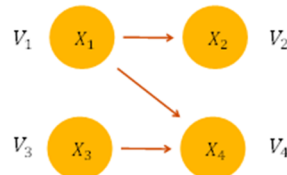


では、それから、観察データ。今、元々すでにあるデータから変数間の因果関係を捉える、「因果探索」と呼ばれる手法をご紹介します。特にベイジアンネットワークと呼ばれる、割と広く使われている手法もございますので、それを中心にご紹介したいと思っています。それらを用いて、どのように、今お話ししていた「因果効果」と呼ばれる、介入による効果を測定していくのかということについて、ご説明したいと思っています。

DAGと因果関係 ~記号の定義~

確率変数間に存在する因果関係を表すため、グラフによる表現を導入する

対象	記号	説明
ノード	V_i	グラフの頂点。確率変数 X_i が対応
親ノードの集合	$\pi(i)$	ノード V_i に向かうエッジの始点となるノード(親ノード)の集合



例:

$$\begin{aligned}\pi(2) &= \{1\} \\ \pi(4) &= \{1,3\}\end{aligned}$$

以下、ノード V_i と確率変数 X_i を同一視した表現を行う

PwC

9

では、ベイジアンネットワークの要素を定義するために、DAGと呼ばれるグラフと、変数、どのような変数が出てくるのか、変数間の因果関係をどう表現するのかというところをご説明します。

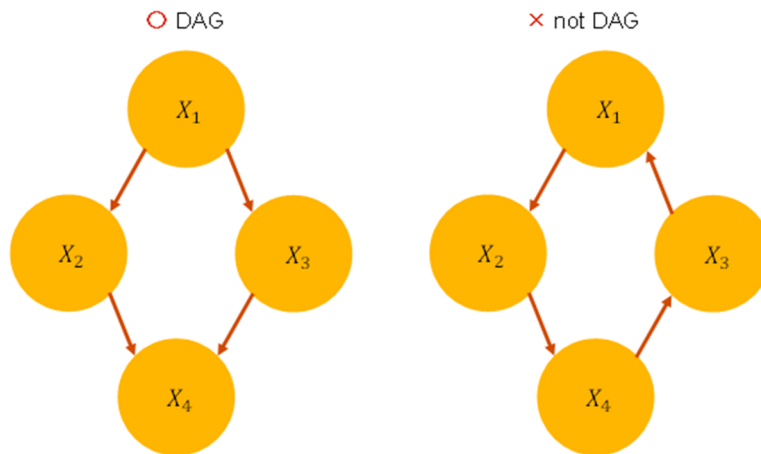
まず、先ほどからちらちらと絵には出しているのですが、この確率変数間の因果関係を表現するために、こういった丸と線でつながれているグラフを導入します。この丸のことを、頂点という意味で「ノード」という言葉で表現します。それぞれのノードは絵的には丸でして、あともう1個「エッジ」という言葉を使います。エッジとは、ノードとノードをつないでいる線の矢印の関係とします。ノードとノードの間を矢印でつないで、矢印の元から先に因果関係があるといったことを表現したいと思います。

特に数学的な記号については、Vに添え字で、ノードを V_i として書いて、そのノードの上に確率変数が対応しているというような表現もします。だから、確率変数とノードを同一視したような表現もするのですが、混乱せずに、あえてそのような表現をしているというところにご注意ください。

それから、「親ノード」という言葉も使うのですが、それを i というノードに対して、親として πi というような表現をします。例えば、今、ここのグラフを見ていただくと、2のところの親は1です、あるいは4のところの親は1と3ですといった、このエッジでつながる関係の親のことを親ノードといっています。

DAGと因果関係 ~DAGの例~

DAG (Directed Acyclic Diagram)を、循環しない有向グラフとして定める



PwC

10

本日の重要なパーツの1個の「DAG」と呼ばれているものですが、こちらは、グラフの内をぐるりと巡回しない。因果関係に沿っていても、循環しないようなグラフと定義します。例えばこの左側のは、エッジに沿っていても、ここに行かないようにぐると回らないのですけれども、こちらはぐると回ってしまうので、このようなものは相手にしないということですね。

DAGと因果関係 ~条件付き独立性~

以降、DAGによって表現できる同時分布の確率を考える

例: $\vec{X} = (X_1, X_2, X_3, X_4)$

一般に成立する同時分布確率表現:

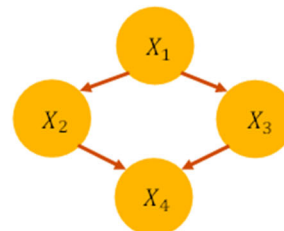
$$p(\vec{X}) = \prod_{i=1}^4 p(X_i | X_1, \dots, X_{i-1})$$

$$= p(X_1)p(X_2|X_1)p(X_3|X_1, X_2)p(X_4|X_1, X_2, X_3)$$

DAGによる因数分解表現:

$$p(\vec{X}) = \prod_{i=1}^4 p(X_i | \pi(i))$$

$$= p(X_1)p(X_2|X_1)p(X_3|X_1)p(X_4|X_2, X_3)$$



PwC

11

では、次ですね。DAGというもので、では、何がいいのかというところですが、これで、本日表現したい確率変数の同時分布を若干制限していきます。一般に、例えば、今、四つの確率変数が並んでいたとして、同時分布を分解していくということが一般的にはできます。こちらはアクチュアリー試験の数学などでも出てきたと思うのですが、1個前の変数で条件づけをしていくことで、それぞれ条件付き確率の積になっていきますが、今、DAGというものの上乘っている確率変数だとういった表現になるかというと、それぞれがX1、X2で条件がついているものがX1だけになったり、ここの1、2、3で条件が

ついているものが2、3だけになるなど、そういったものをDAGと一緒に考えるということが、今日、重要になってきます。

ベイズアンネットワークの定義

同時分布確率 $p(\vec{X})$ が、DAG構造 G と親ノード $\pi(i)$ における条件付確率表(CPT) θ の組を用いて積表示できるものをベイズアンネットワークと定める

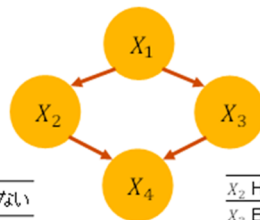
例:

$$\pi(2) = \{1\}, \pi(3) = \{1\}, \pi(4) = \{2,3\}$$

$$p(\vec{X}, G, \theta) = \prod_i p(X_i | \pi(i))$$

$$= p(X_1) p(X_2 | X_1) p(X_3 | X_1) p(X_4 | X_2, X_3)$$

X_1 飲酒		
飲む	θ_{11}^1	40%
飲まない	θ_{21}^1	60%



X_1 飲酒	飲む	飲まない
X_3 BMI		
低	θ_{11}^3	70%
高	θ_{21}^3	30%

X_1 飲酒	飲む	飲まない
X_2 HbA1c		
低	θ_{11}^2	90%
高	θ_{21}^2	10%

X_2 HbA1c	低	高
X_3 BMI	低	高
X_4 入院発生		
なし	θ_{11}^4	99%
あり	θ_{21}^4	1%

PWC

12

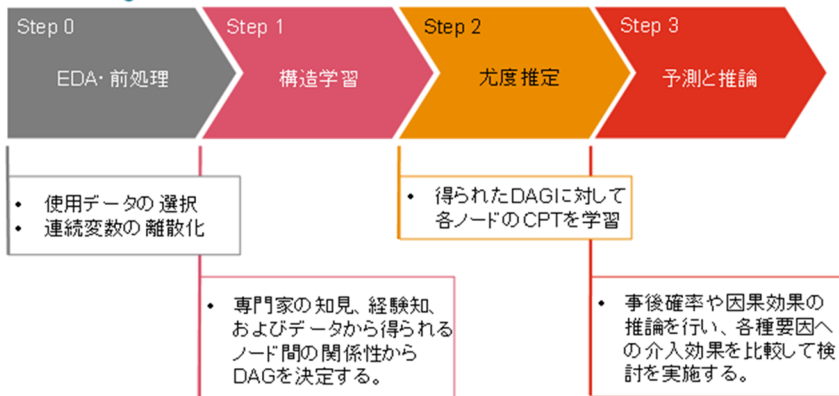
ここまで来ると、今日の主題の「ベイズアンネットワークとは何ですか」ということを定義できるのですが、この同時分布ですね。同時確率分布と、DAGの構造と、それから親ノードにおける条件付き確率表、Conditional Probability Table、CPTといたりしますが、そこに乗っているパラメータの組を用いて積表示ができるものをベイズアンネットワークとして定めます。例えば、先ほどから書いている、この四つの変数のDAGでいきますと、先ほどお話ししていた条件付き確率の積として、比較的単純になりますが、そのノードの上にこういった確率分布の表が載っていると思ってください。

もっと身近な例として、それぞれの確率変数が、飲酒、HbA1c、BMI、入院発生のアウトカムと考えたときに、飲酒のノードでは、飲む・飲まないというのが二つに分けられて、40%や60%というように乗っています。X2のノードの上では、こういったテーブルが乗っていて、HbA1c(血糖値の指標)、これでは親のX1で条件がついていて、親のX1の条件ごとにこういった確率が決まっているといった感じのものを、想像していただければいいかと思います。

では、ここまでベイズアンネットワークが定義できましたので、これを使ってどう因果探索をやっていくのか、因果推論をしていくのかといったところを、堤さんから紹介してもらいます。

因果探索・推論の手順

BNを推定するステップは大きく分けてDAG推定とCPT推定の二つに分けられる。Pythonライブラリ  pgmpyを用いる



PwC

13

堤 板倉さん、ありがとうございます。これから、私、堤の方から、ベイジアンネットワークの活用方法について、ステップごとに紹介させていただきます。ベイジアンネットワークの活用の手順につきましては、こちらのページに記載している4ステップからなっております。

まず、ステップ0として、EDAや前処理を行ってまいります。欠損値補完のような通常のEDAに加えて、ベイジアンネットワークでは変数を離散的なものとして扱うため、データの離散化などが必要となっております。

ステップ1として、構造学習がございます。ステップ1ではDAG構造を決定いたします。DAGの構造の決定には、先行研究からの知見や専門家の判断、個人的な経験、また、それらに加えて、データセットから学習できるノード間の関係性もこの構造学習のステップでは使用いたします。

その次のステップ2としましては、尤度推計としてCPTのパラメータを学習いたします。CPTの学習には、最尤推定を本日は使用いたしました。

ステップ2まででベイジアンネットワークの学習が完了いたしまして、各種確率推論ができるようになります。ベイジアンネットワークにおける確率推論は事後分布、つまり、条件付き確率のことですね。条件付き確率や、このあと説明する、因果効果といわれる事後確率の一種のようなものも推定することができるようになります。因果効果は、先ほどご説明した、交絡因子からの疑似相関の影響を除いた事後確率として計算することができまして、施策の比較検討や意思決定に使用することができます。

Step1: 使用する知識源

Step 0 EDA前処理 Step 1 構造学習 Step 2 尤度推定 Step 3 予測と検証

DAGはいくつかの知識を組み合わせて作成する(*1)。

先行研究

- 先行研究で知られている知見をノード間の関係性として反映させる。

専門知識・ 個人的な経験

- 論文に限らず、専門家の知見や、個人的な経験も知識源になりうる。
- 個人の経験の使用には個人の思い込みが混入しないように注意。

因果構造 探索

- データセットからノード間の因果関係を自動的に学習すること。
- 今回はスコアベース学習と呼ばれる方法を使用。

(*1) 『Pythonライブラリによる因果推論・因果探索(概念と実践)因果機械学習の鍵を解く』(Aleksander Molak 著 / 株式会社クイープ 訳)
(*2) https://www.istage.ist.go.jp/article/s.ekatsueisei/1957/326/32_6_280/pdf/-charia (大阪市立環境科学研究所)

PwC

14

この次のページで、各ステップについて説明いたします。まず、ベイジアンネットワークを構築するために、DAGの構造学習を行います。DAGの構造は、こちらに記載している三つの知識を組み合わせて決定いたします。

まず、一つめの知識として、先行研究で知られている知見をDAGに反映するといったことが挙げられます。

そして、二つめの知識として、専門家や各個人が持っている経験的な知識も、DAGを構築するための知識源に十分になります。例えば、自社の被保険者集団の特徴として、商品ごとに入院発生率が違っているというようなことが知られていれば、その知識をDAGに反映することも考えられます。ただし、こういった知識は、個人の思い込みなどから誤ったモデルになる可能性があります。注意して使用する必要がございます。

最後に、データ自体からDAGの構造を決定することも可能になります。これを因果構造探索と呼んでおります。今回はスコアベース学習と呼ばれる手法を用いて、因果構造探索を行いました。スコアベース学習の説明はこのあと行いますが、その前にCPTを表現するパラメータの定義と、それを用いた尤度関数を紹介させていただきます。

Step1&2: CPTのパラメータ表示

Step 0 因果処理
Step 1 確率学習
Step 2 尤度推定
Step 3 予測と検証

BNでは、変数を離散化し、親変数の状態で条件付けした多項分布でCPTを表現する。
その際の多項分布のパラメータ集合 θ を以下で定義する:

- $\theta = \{\theta_{k|j}^i | 1 \leq i \leq n, 1 \leq j \leq q_i, 1 \leq k \leq r_i\}$
- $q_i: \pi(i)$ の取りうる状態数
 - $r_i: X_i$ の取りうる状態数
 - $\theta_{k|j}^i: \pi(i)$ が j 番目の状態の下で X_i が k 番目の値をとるパラメータ ($\sum_{k=1}^{r_i} \theta_{k|j}^i = 1$)

例) X_4 の場合:

$q_4 = 4$

X_2 HbA1c	1	1	2	2
X_3 BMI	1	2	1	2
X_4 入院発生	1 (なし)	$\theta_{1 1}^4$	$\theta_{1 2}^4$	$\theta_{1 3}^4$
	2 (あり)	$\theta_{2 1}^4$	$\theta_{2 2}^4$	$\theta_{2 3}^4$

$r_4 = 2$

尤度関数と最尤推定値は以下で与えられる:

- 尤度関数: $L(\theta, D) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{(\sum_{k=1}^{r_i} N_{k|j}^i)!}{\prod_{k=1}^{r_i} N_{k|j}^i!} \prod_{k=1}^{r_i} (\theta_{k|j}^i)^{N_{k|j}^i}$
- 最尤推定量: $\hat{\theta}_{k|j}^i = \frac{N_{k|j}^i}{\sum_{k=1}^{r_i} N_{k|j}^i}$
 - $N_{k|j}^i$: データセット D において、 $\pi(i)$ が j 番目の状態の下で X_i が k 番目の値をとったデータ数

PwC

15

CPTを表現するためには、こちらに記載している θ のような、確率を表すパラメータの集合により表現されます。定義だけだとイメージがわきづらいかと思うので、先ほど紹介していた、入院発生のノードを例として紹介させていただきます。入院ノードの親変数にはHbA1cとBMIの二つがございまして、これらの二つの状態の組み合わせとして、入院発生のノードの親変数の状態としては四つを取り入れます。そういった各変数の状態ごとに条件づけられたときの入院発生の有無が、 θ によって多項分布として表されているといったものをイメージしていただければと思います。

このようなパラメータ集合 θ を用いた尤度関数は、こちらに記載している式で表現されます。こちらの式の中の n という記号につきましては、対応するCPTの要素に対して、データセットの中で、この要素が幾つ観察されたかを表す変数になっております。この尤度関数の形については、各多項分布の尤度関数の積のような形になっているとご理解いただければと思います。

そして、この尤度関数から得られる最尤推定量は、こちらに記載しているものになっております。多項分布の最尤推定量は、よく知られているように、ラグランジュの未定乗数法を用いて求めることができます。これらの尤度関数と最尤推定量を用いて使うスコアベース学習と尤度推定の説明をしていきます。

Step1: スコアベース学習

Step 0 因果処理 Step 1 構造学習 Step 2 尤度推定 Step 3 予測と検証

- 各DAGのスコアを計算し、より高いスコアのDAGを採用する手法。
- 候補となるDAG構造は数が多く、実際にはヒルクライムサーチという探索方法で局所的な最大値を探索した。(Appendix p.47 参照)

BICスコア:

$$\text{BIC}(G, D) = \sum_{i=1}^n \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{k|j}^i \ln \left(\frac{N_{k|j}^i}{\sum_{k=1}^{r_i} N_{k|j}^i} \right) - \frac{1}{2} \ln(N) \sum_{i=1}^n (r_i - 1) q_i$$

最大尤度 正則化項

- G : DAG
- D : データセット

計算例 (*):



スコア最大のDAGを採用

PwC

(*) 1万サンプルのデータを生成して計算した。

16

まず、スコアベース学習の説明をさせていただきます。こちらはどのようなものかといいますと、まず、与えられたDAGの候補に対して、そのDAGがデータセットをどの程度よく表現しているかを表すスコアを計算して、そのスコアが最も高いDAGを採用するといったものになっています。今回使用したスコアですが、BICスコアを使用しております。BICスコアをこちらに記載しているというような算式になっておりますけれども、まず、第1項目が、先ほど紹介した尤度関数の対数の最大値になっております。そして、2項目が、DAGのパラメータ数に対してデータセットの数の対数を掛けたものとして、正則化項となっております。

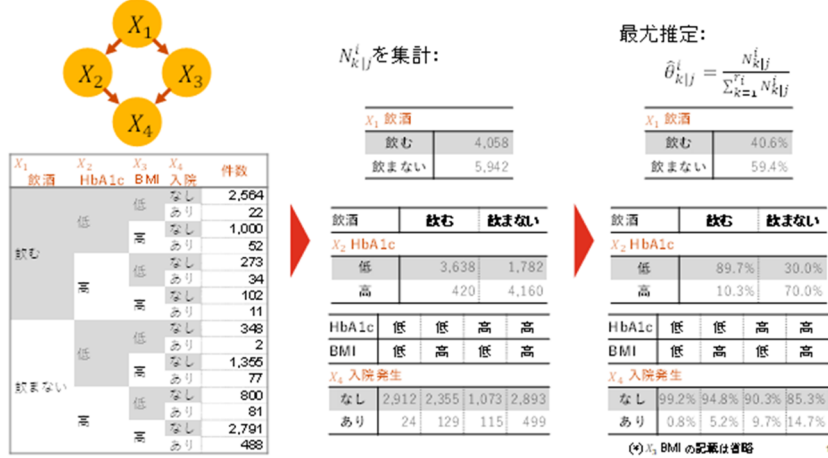
ページの下半分に、サンプルデータを使用して計算したスコアの例を記載しております。三つのDAG候補の中では、真ん中のDAGが一番スコアが高いものとなっております。スコアベース学習では、こちらの真ん中のDAGが採用されるといったようなものになっています。一方、一番右のDAGについてはエッジが多くて、1個目の最大尤度の項は大きくなるのですが、正則化項の影響によって過学習が抑えられているといったような様子が見て取れるかと思います。

Step2: 尤度推定

Step 0 ED/A 前処理 Step 1 構造学習 Step 2 尤度推定 Step 3 予測と推論

各ノードの親ノードの状態条件付けた条件付確率を多項分布とみなし、そのパラメータを最尤推定法により推定する。

サンプルデータを用いた例:



そのあと、構造学習により DAG の構造を決定したあとは、CPT のパラメータ θ をデータから推定いたします。パラメータ θ の推定には、ベイズ的な手法で学習する方法もありますが、今回は、先ほど紹介した最尤推定を用いて学習しております。こちらのページの左下には、サンプルデータを用いて学習、最尤推定を行った例が記載されていますけれども、先ほど最尤推定量は紹介させていただきましたので、本日は時間の関係で、こちらの紹介は少し省略させていただきます。

Step3: 因果効果推定

Step 0 ED/A 前処理 Step 1 構造学習 Step 2 尤度推定 Step 3 予測と推論

DAG を用いると X (例: 生活習慣) に介入したときの Y (例: 入院) に与える因果効果を推定することができる:

因果効果の推定式^(*)

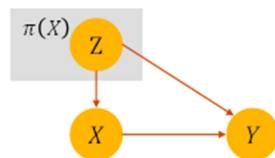
$$p(Y = y | \text{do}(X = x)) = \sum_{j=1}^{q_i} p(Y = y | X = x, \pi(X) = \pi_j) \times p(\pi(X) = \pi_j)$$

• π_j ($1 \leq j \leq q_i$): $\pi(X)$ の取りうる状態

事後確率と因果効果

通常の事後確率 (伝統的保険数理):

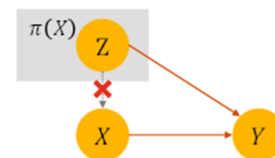
$$p(Y, Z | X) = p(Y | Z, X) p(Z | X)$$



事後確率 $p(Y | X)$ は $p(Z | X)$ を通じて交絡因子からの影響を受ける。

因果効果:

$$p(Y, Z | \text{do}(X)) = p(Y | Z, X) p(Z)$$



X と Z を独立にすることで交絡因子の影響を排除。

(*) 因果的推論モデル・推論・推論 (Judea Pearl 著 / 黒木孝 訳)

最後は、CPT の最尤推定まで完了しますと、事後確率や因果効果の推定が行えるようになります。ここでは因果効果の推定式を紹介いたします。こちらに記載している式が、変数 X から Y への因果効果の推定式となっておりまして、算式中の do という記号は、単なる事後確率ではなくて、因果効果を推定するといったことを表しております。こちらですが、詳しい算式は参考文献を参照することとさせていただきます。こちらの因果効果も事後確率の一種なのですが、通常の事後確率との違いを、例を用いて少し紹介させて

いただきます。

まず、通常の事後確率は、従来の保険数理において、年齢別や性別の死亡率を計算するものに相当しておりますが、この事後確率には、冒頭で紹介したような疑似相関の影響を受けてしまい、施策の比較や意思決定には適さないといった注意点がございます。とても極端な例ではございますが、先ほどの冒頭で紹介した、アイスの売上ごとの熱中症患者の事後分布を見て、アイスの売上が熱中症患者を増加させていると判断して、熱中症患者を下げるためにアイスの販売を制限させるといった施策を考えます。しかし、この場合は、アイスの売上を減少させたとしても、熱中症患者は下がることはなく、もちろん気温を下げないと熱中症患者は減少するわけではないので、誤った意思決定をしてしまうことになってしまいます。現実の意思決定の場面においても、これほど極端な例ではないにしろ、疑似相関に気づかず、誤った判断をしてしまう可能性は十分に考えられます。

一方で、因果効果で計算しているものは、右下のように、交絡因子ZからXへのエッジを削除して、XとZを独立させた確率分布での事後確率を考えているものになっております。このような確率分布を考えることで、気温のような交絡因子による疑似相関の影響を排除して、Xを変更したことによるYへの直接的な効果を見ているといったものをイメージしていただければと思います。このような因果効果は、施策の比較や意思決定に使用することができるようになります。

この因果効果は、伝統的な保険数理におけるような単なるデータ集計では得られないものとなっております。変数間の関係性を学習することで、初めて交絡因子の特定と、それに基づいた因果効果の推定ができるようになります。このような因果効果の推定ができることが、因果探索の大きなメリットと感じております。

ベイジアンネットワークの数理的なところは以上になりまして、このあとはベイジアンネットワークの活用事例に入らせていただきます。

板倉さん、お願いいたします。

アジェンダ

- 1 課題設定
- 2 因果探索と
ベイジアンネットワーク
- 3 ベイジアンネットワークの
DeSCデータへの適用例
- 4 本日のまとめ
- 5 Appendix

PwC



板倉 はい、ありがとうございます。それでは、ベイジアンネットワークおよび因果探索の基本的な事項の

ご説明は以上になりまして、これからDeSCヘルスケアさんのデータセットを使ったときの適用例をご紹介します。

ベイズネットワークの適用例

脳血管疾患・心疾患入院に関する因果探索・因果推論以下のテーマで紹介する。



■ ケース1「リスク評価」

- アウトカムに対する直接のリスクファクターは何か
- 通常の加算方式のリスク評価に対し、因果関係を考慮したリアルワールドベースの評価による保守性の除去



■ ケース2「介入施策比較」

- 2CIIによる入院発生率を低減するために効果的な生活習慣の改善項目は何か
- 生活習慣改善の継続を介入とみなした因果効果の推定

(*) DeSCヘルスケア社提供のデータセットを用いて分析を行った。受検データの詳細はAppendix p37を参照。
以降、脳血管疾患および心疾患をまとめて2CII (Critical Illness) と呼ぶ。

PwC 20

では、まず、一つ目の方ですけれども、先ほどお話ししていました、バリューチェーンの右上と右下のところなのですけれども、一つ目は右上の方で、引受リスク評価の場面での活用といったものを若干簡単なトイモデルで検証します。こちらでは、両者ともに、脳血管疾患および心疾患入院といったものをアウトカムとして考えていきます。どちらも今の手法を使っていくのですが、まず一つ目の方は簡単な手法でということで、どのように挙動しているのかを理解していただくうえで、従来のリスク評価の手法と、今回のリアルワールドの因果関係を評価したようなリアルワールド評価を比較していきます。

もう一つのケース2の方では、介入施策の比較を行います。ここでは、喫煙や飲酒などといったような生活習慣のファクターと、健康診断の項目および先ほどのアウトカムと、3種の変数間の因果関係を観察データであるDeSCヘルスケア社のデータを使ってDAGを決定して、その得られたDAGの上で生活習慣への介入を行って、因果効果を推定していくといったことを行います。

DeSCヘルスケアが提供するデータ



DeSCヘルスケアは以下3種類の保険制度にかかるレセプトデータを匿名加工情報として提供*しています。

- 健保（健康保険組合）
- 国保（国民健康保険）
- 後期高齢（後期高齢者医療制度）

保険制度	保険者	被保険者
健康保険組合	企業	企業の従業員と家族 基本的に64歳以下
国民健康保険	市区町村	自営業者等企業に属さない者 74歳以下
後期高齢者医療制度	都道府県	基本的に75歳以上 (65～74歳は一定の障がいを含む)

(*) DeSCヘルスケアが提供するデータの最新情報はDeSCヘルスケア会員制メディア
(<https://www.desco-life-actuary.jp/>)にて確認できます

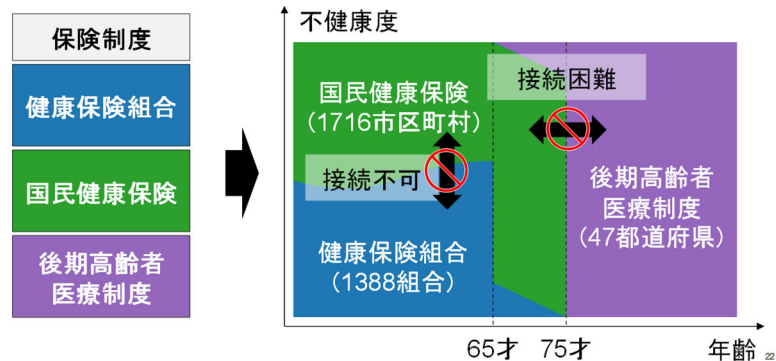
21

今回、分析データとしてお借りしているDeSCヘルスケア社のデータセットの特徴ですが、保険者制度別のレセプトデータを匿名加工したものとなっています。対象となる保険制度は、健保組合、国保、後期高齢の三つになっています。個人IDでこれをひもづけることで、各種レセプトや健診・問診データを使った分析、もしくは年度をまたいだ分析をすることが可能です。DeSCヘルスケア社のデータに関する最新情報等に関しては、こちらのURLから確認いただければと思っています。

保険制度ごとのデータの特徴



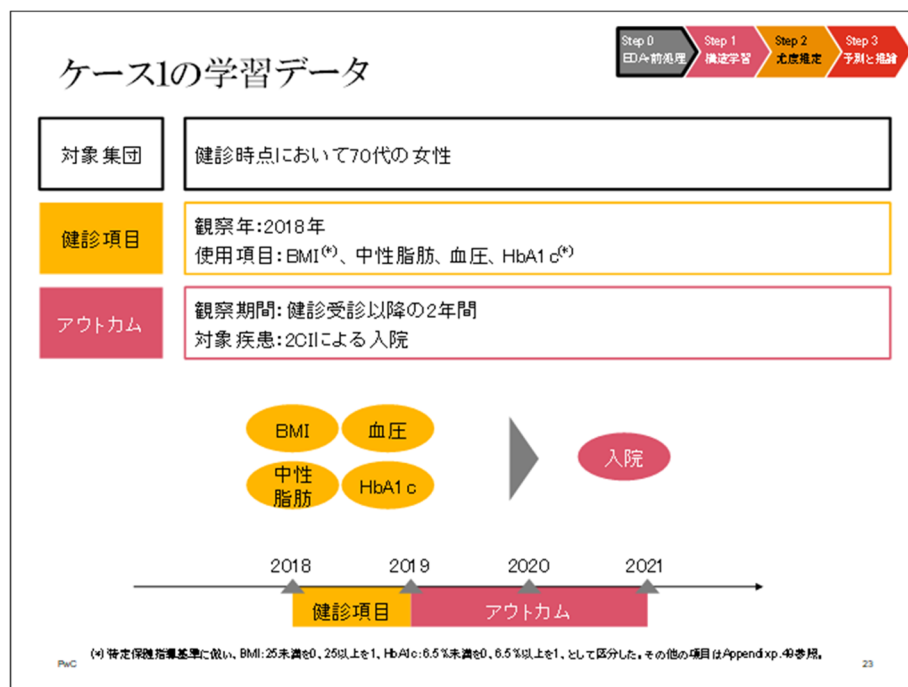
- 3つの保険制度を年齢と不健康度（不健康な人が含まれる割合）でマッピングすると以下のように示すことができます
- 健保データは(1)健康バイアスが強く(2)死亡分析ができず(3)64歳までしか分析が困難でした
- 国保・後期高齢データの利用により(1)不健康者の分析(2)死亡分析(3)高齢分析ができるようになります



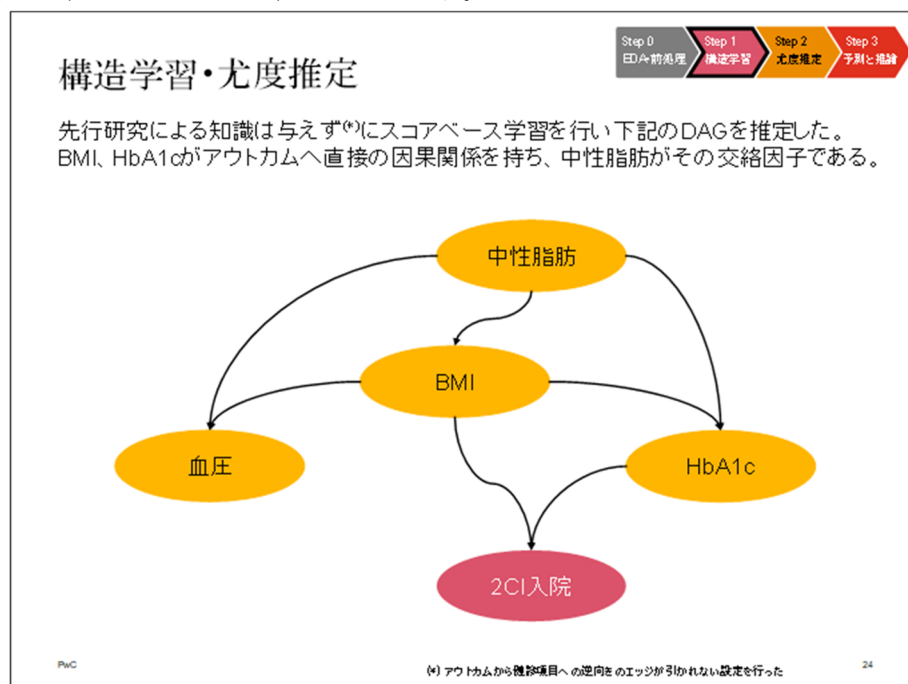
次に、このデータセットならではの特徴なのですが、まず前提として、健保データでは死亡分析や後期高齢者部分の分析ができないという問題点はあるのですが、そのうえで、国保や後期高齢のデータを用いて死亡や長期の観察ができるといったところが可能になってきます。

このあと改めてご紹介しますが、今日は高齢層における健診データを用いた分析事例になっています。こ

ういった分析ができるということも、健診データをひもづけられる高齢者のデータが多いといった、DeSCヘルスケア社さんの特徴になってくるかと思っています。詳しいお話は、DeSCヘルスケア社の方にお伺いしていただければと思います。

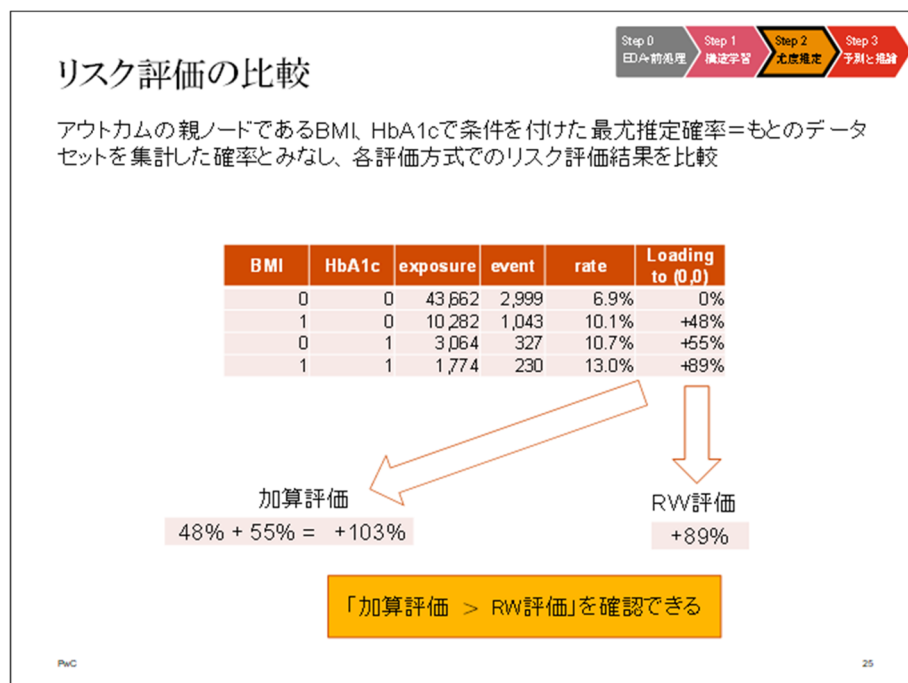


ケース1に入っていくのですが、こちらで使用したデータの対象の集団は、2018年に健診を受診した時点で70代である女性としています。使用する健診項目は、比較的皆様になじみのあるBMIや中性脂肪、血圧、HbA1c(血糖値の指標)としました。そして、アウトカムは、健診の受診時点から2年以内に発生する2C I、先ほどの脳血管疾患、心疾患による入院としています。



これらの健診項目、アウトカムに対して、先ほどのスコアベースの学習によって、DeSCデータから因果関係を探したものがこちらに書いているDAGになります。このとき、特に入院から健診項目にエッジを引くということが何もしないと起きてしまうのですが、そういったことが起きないように学習をさせてい

ます。この結果、注目していただきたいことは、BMI、HbA1c がアウトカムに直接影響を与えるファクターであるといったこと、および、中性脂肪がそれぞれのファクターの交絡因子になっているということが、分かるかと思います。これを用いてBMIとHbA1cが直接のリスクファクターということが分かりますので、従来の保険引受リスク評価がどうなっているのか、今回のリアルワールド評価とは何が違うのかということをご説明します。

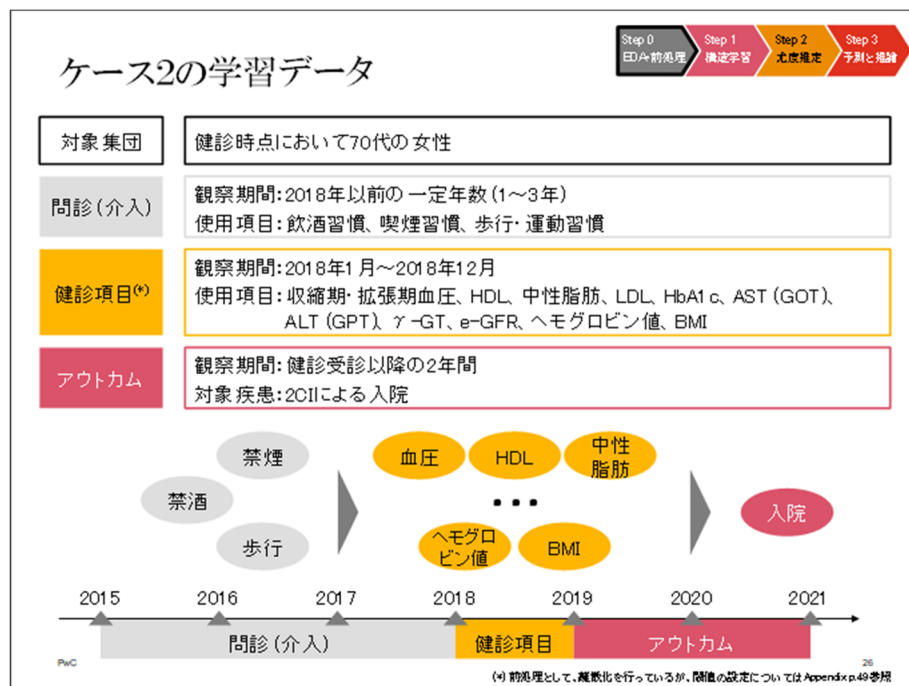


こちらはデータセットを抽出してきて、今、BMIとHbA1cの条件を離散化させて作っているのですけれども、0か1かと。イメージとしては0は標準体で、1は条件体以上といったものを想像していただければいいかと思います。詳しい条件はAppendixにつけておりますが、これは厚生労働省さんの指標を用いて離散化しています。

上側には、BMIとHbA1cそれぞれの区分ごとに、対象者の件数や入院発生件数、入院発生率を計算しています。特にこの(0,0)といったものは、先ほどお伝えした標準体をイメージしていただければいいかと思います。こちらのLoading to (0,0)に関しては、標準体に対する評点ですね。例えば、(1,0)に対しては、10.1% ÷ 6.9%で、プラス48%と書いてあると思っていただければいいです。この表を基に、従来の加点方式をするとうなるかというと、例えばですけれども、(1,1)の人の評価は、この(1,0)の人と、(0,1)の人の評価を足し上げて101%だと評価をするというものが、よくやるアンダーライティングの手法です。

一方で、今、データから学習させた結果で分かることは、プラス89%を示しているというようになっているので、この加算評価、従来の評価と実際にリアルワールドで評価した評価は、この差分だけ違うといった考え方で保険リスクを評価すれば、まっとうなといいますか、保守性が削除できるのではないかなというように考えられます。

では、イメージはこれで持っていただけるかと思うので、ケース2の方は堤さんにご説明してもらおうと思います。では、よろしくお願いします。



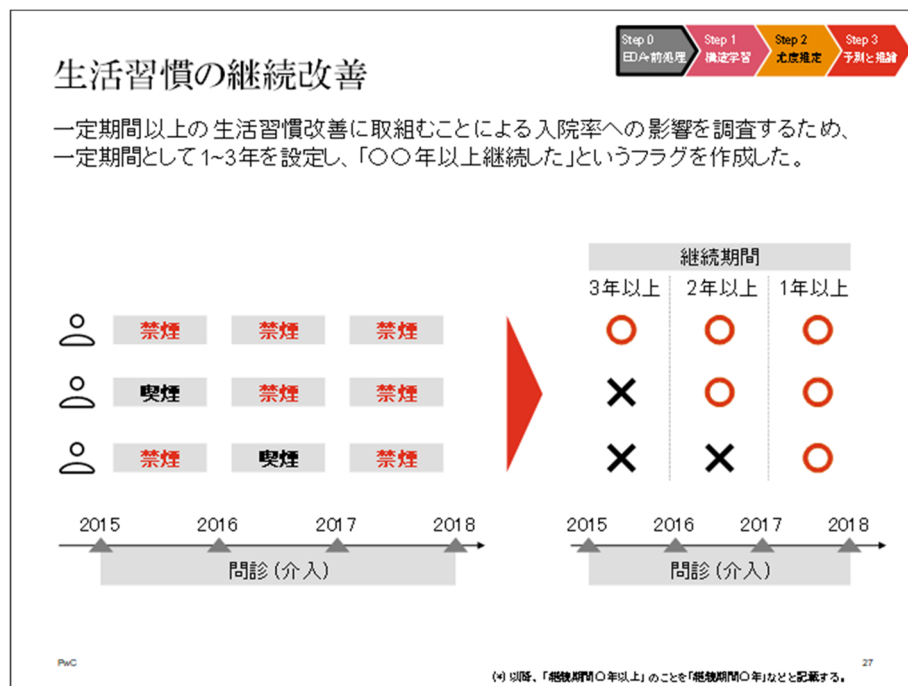
堤 はい。板倉さん、ありがとうございました。

これから、私の方からケース2の紹介をさせていただきます。ケース2では、喫煙習慣などの各生活習慣に対して介入、つまり改善行動を促した場合に、どの生活習慣への効果が最もあるのか、また、血圧等の体況ごとに因果効果は異なるのかといった施策比較の場面を想定したものとなっております。

まず、学習に使用したデータを説明いたします。対象集団とアウトカムについては、ケース1と同じになっておりますので、少し省略させていただきます。

次に介入項目についてですが、こちらは問診項目を使用しております。飲酒と喫煙、あと、歩行習慣に関する問診項目を採用いたしました。これらの項目を2018年までの3年間観察して、生活習慣を長期間改善することによる影響を今回は見ております。次のページで、こちらについては改めて説明いたします。

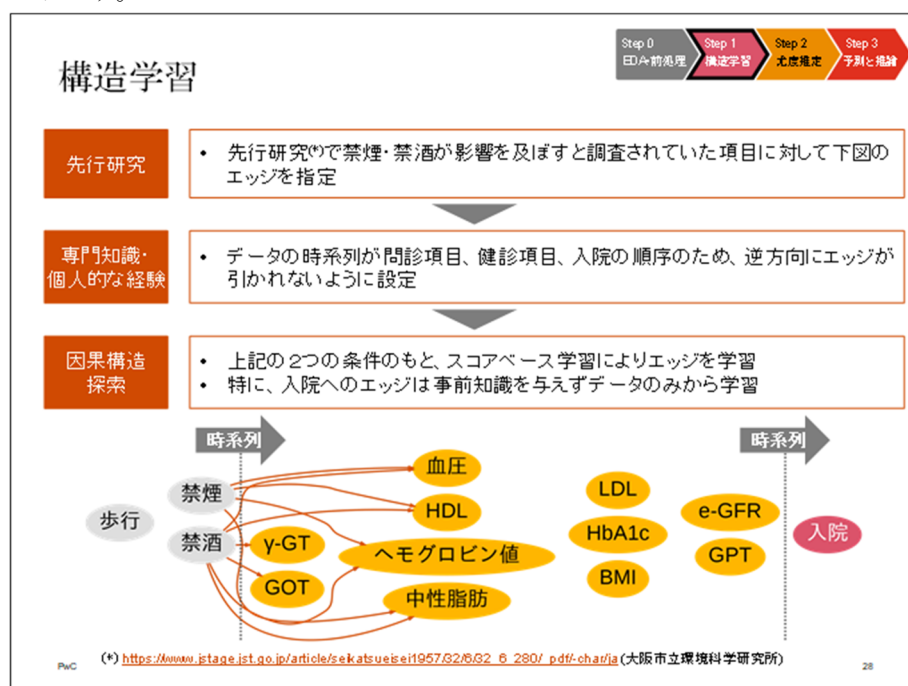
次に健診項目ですが、2018年に受診した健康診断の数値を採用しています。項目としては、こちらに掲載している12項目を使用しております。また、こちらの項目については、前処理の段階で離散化を行っておりますが、離散化の詳細についてはAppendixを参照していただきます。これらのデータの時系列については、下に記載しているとおり、まず介入である問診項目を観察して、その次に健診項目、最後にアウトカムを観察したというような時系列になっております。



先ほど生活習慣を改善したことによる影響を観察したと説明いたしましたが、具体的にはこちらのページにあるように、2018年の直近3年間、継続して改善を行っていたかを観察しています。例えば一番上の被保険者は、3年間すべて禁煙を行っていた被保険者となっておりまして、こういった被保険者は、2018年時点で3年以上の禁煙者でありますし、2年以上、1年以上の禁煙者としても分類して、分析を行っております。

2人目の被保険者についても、同様の処理を行っております。

3人目の被保険者につきましては、途中で喫煙を行ってしまっています。この場合は、2018年の直近1年間のみ喫煙、禁煙を継続していた被保険者として分析しておりまして、2年以上禁煙を行っていない被保険者として分類しております。以降は、3年以上から1年以上まで、それぞれの継続期間に分けて分析した結果を紹介してまいります。

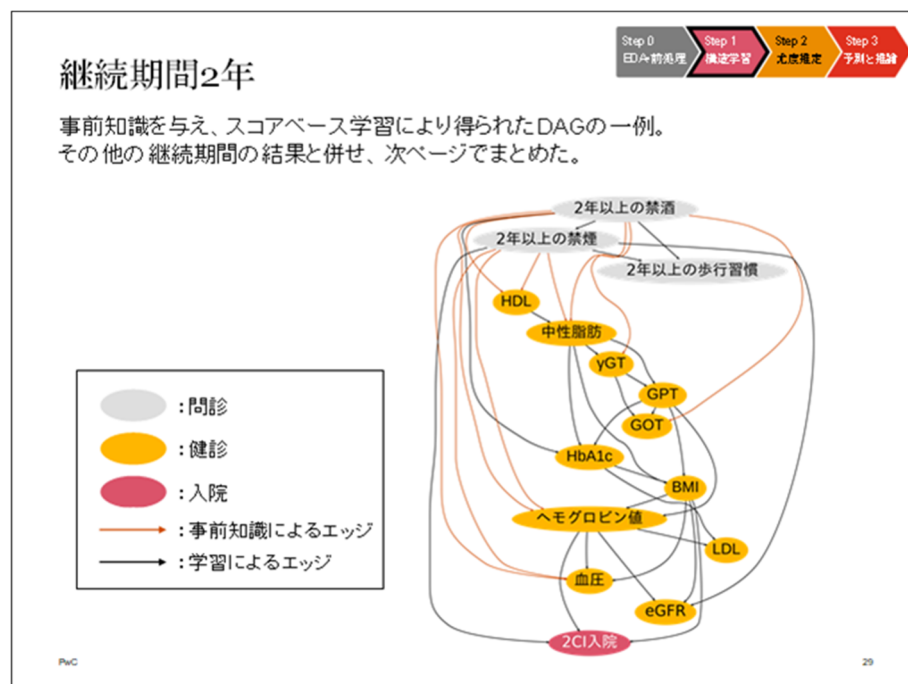


次に、ケース2の構造学習において使用した知識を紹介いたします。まず、先行研究として採用した知識

については、喫煙および飲酒が健診値に与える影響を調査した研究を参考にいたしました。その研究で関係性が認められていた項目について、エッジを指定しています。そのエッジが、下の図に記しているオレンジ色のエッジとなっております。

次に、データを観察した時系列が、問診、健診、最後に入院という順番となっているため、各変数の因果関係もこの時系列に従うということが自然と考えられますので、エッジがこの時系列に反して引かれないように制限を行いました。

最後に、これらの条件のもとで、スコアベース学習によってその他のエッジを学習して、DAGを構築しております。スコアベース学習では、これらの先行研究の知識の取り込みや各種の制限を比較的織り込みやすい手法となっております。実務で応用を検討する際に、非常に使い勝手がよいものと考えております。それでは、学習したDAGの結果を紹介してまいります。



こちらは、継続期間が2年のDAGの学習結果となっております。はじめに図の見方ですけれども、このオレンジ色のエッジについては、先ほど紹介した先行研究から設定したエッジとなっておりまして、黒色のエッジがスコアベース学習によって学習されたエッジとなっており、このように、実際のデータから変数間の関係性が学習されているといったようなことが見て取れるかと思います。

次のページで、まとめた結果を紹介するのですが、まずこちらのページで注目していただきたい点が三つの生活習慣の関係性でございます。まず、禁酒が禁煙と歩行習慣の親ノードになっており、さらに禁煙についても歩行習慣の親ノードになっている点です。特に歩行習慣は、禁酒と禁煙という二つの交絡因子からの影響を受けているといったような結果になります。実は、この関係性は、継続期間1年および3年においても同様の結果が得られておりまして、生活習慣の間の関係性が示唆されたような結果と考えております。次のページで、1年と3年の結果も併せて紹介いたします。

継続期間別比較 2018年

Step 0
E/A 前処理

Step 1
構造学習

Step 2
尤度推定

Step 3
予測と検証

- Ⅰ 3つの生活習慣は継続期間ごとで同じエッジが引かれていた。
- Ⅱ 長期間喫煙を行っている場合、入院へ直接エッジが引かれていた。
- Ⅲ 歩行習慣は継続2年においてどこにもエッジが伸びていない。

	子													
	歩行習慣	禁煙	禁酒	血圧	OGT	OGT	HDL	HbA1c	LDL	BMI	中位喫煙	eGFR	rCT	ハモロビン
親	歩行習慣									Ⅲ ① ③				
	禁煙	①②③		①②③			①②③				①②③ ②			①②③ Ⅱ ②③
	禁酒	①②③	①②③	①②③	①②③		①②③	①②③			①②③	①②③	①②③	
	血圧													
	OGT					③							①	
	OGT				①②			②③		②③			①	②
	HDL								①	①②				
	HbA1c					①			①②③	②	③			
	LDL													
	BMI			①②③	①		① ③			③	①②③		①②③	①②③
	中位喫煙				①②	③	①②		①②		① ③	②		
	eGFR													① ③
	rCT				②③	②③								
	ハモロビン			②③				①②③			①②③			①②
	201入院													

① 丸数字は継続期間、オレンジ色の文字は指定した事前知識

30

こちらに記載している表がまとめた結果となっております。表の見方ですが、まず、こちらに書いてあるものが親ノードになっております。上の横に書いてあるものが子ノードになっておりまして、対応するセルに記載してある数字が、各継続期間においてエッジが引かれているということを表しております。例えば、禁酒から歩行習慣へのセルを見てみますと、1 から 3 まで記載されておりますので、継続期間 1 から 3 年、すべてにおいて禁酒から歩行習慣へのエッジが引かれていたといった見方をいたします。

先ほど紹介いたしました生活習慣の間の関係性に加えて、今後の事後確率や因果効果の結果を解釈するに当たって必要となる注目すべき点について、二つほど追加で紹介いたします。まず、2 番で記している、禁煙から入院へのエッジが引かれている点です。こちらについては、継続期間 2 年以上で引かれておりまして、より長期継続することによって、入院への影響が大きくなるといったようなことが見て取れます。

3 点目として、歩行習慣についてです。歩行習慣については、継続 2 年目においてどこにもエッジが伸びておらず、その他の期間では BMI のみエッジが伸びておりました。こちらは因果効果を推定する際に大きな影響がございますので、こちら覚えておいていただければと考えております。

以上の 3 点について、このあとの事後確率の推定や因果効果の推定の結果の解釈に関わってきますので、少し頭の片隅に置いていただければと思います。

構造学習のまとめ



結果

- 「生活習慣の間の関係性」、および「生活習慣 (特に喫煙) の改善期間が長くなるほどアウトカムとの因果関係が強くなること」が因果構造探索により示唆された。

留意点

- 結果の安定性・解釈可能性向上のため、各種の知識源から適切な知識を織込むことが重要である。
- 因果構造探索によりデータから得られるエッジは、妥当性を検証することが必要となる。先行研究や、自社の経験などから判断し、モデルの信頼性確保を図る姿勢が求められる。

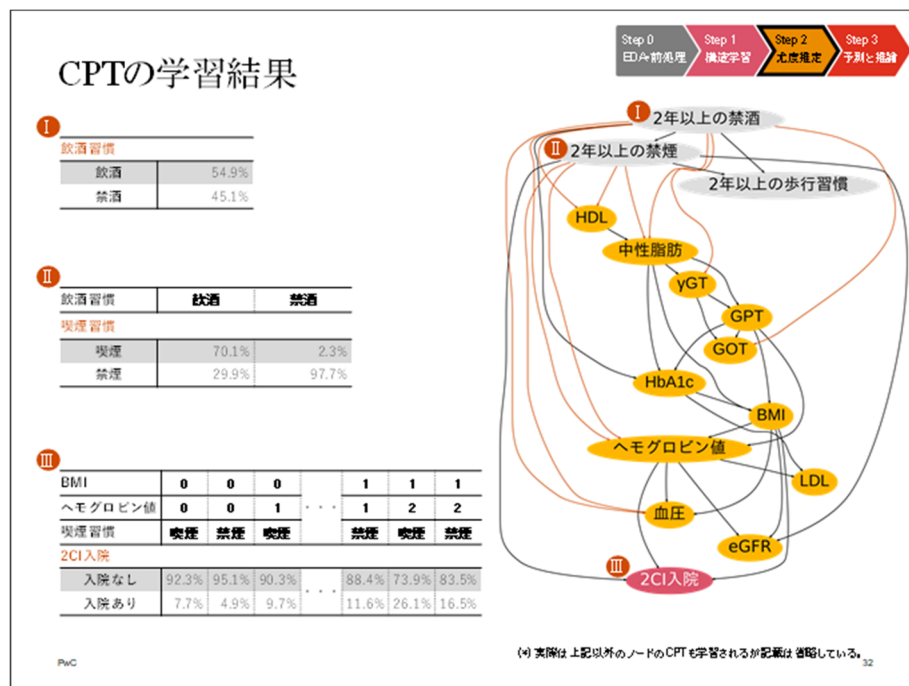
得られたDAGをもとにCPTを学習し、事後確率・因果効果を計算し
各施策の実施による入院発生率への影響を評価する

PwC

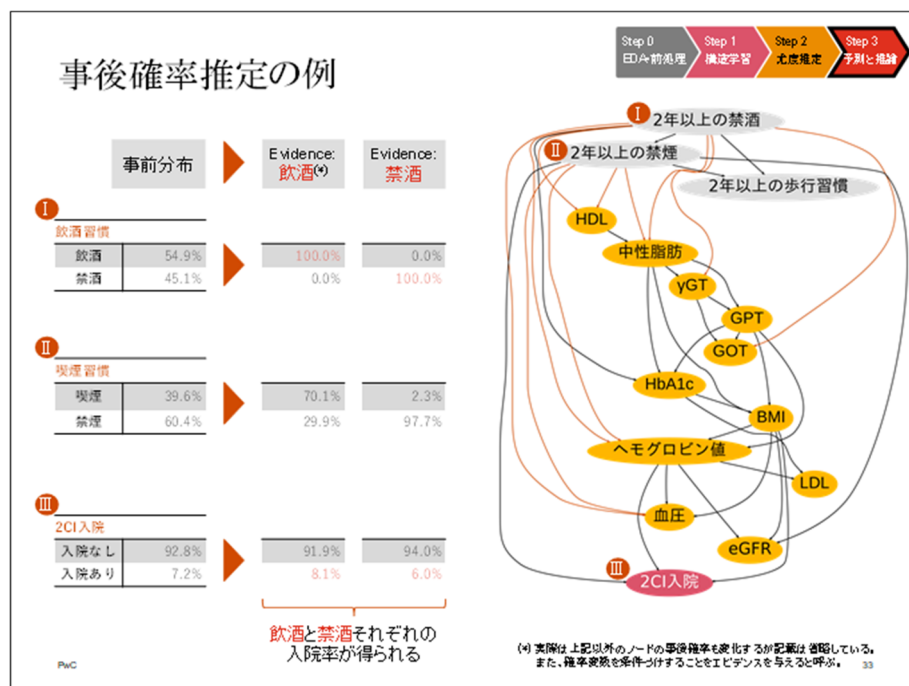
31

こちらのページで、構造学習についての結果と留意点のまとめをいたします。結果については、先ほどのページで紹介した、問診項目間の関係性と生活習慣、特に禁煙ですが、長期間継続することによって、入院との因果関係が強くなるといったことが得られました。実務での応用例の留意点として、まず結果の解釈可能性を向上させていくことが必要と考えております。本日は紹介できていませんが、先行研究からの知識と時系列に関する制限を与えずに、スコアベース学習のみでDAGを学習した場合に、時系列の逆転が発生して解釈が難しい結果となってしまうことも確認しています。こういったことを避けるために、先行研究からの適切な参照などが必要と考えております。

また、因果構造探索により得られたDAG構造は無条件で採用するものではなくて、その結果の妥当性の検証も必要と考えております。得られたDAGにおけるエッジは、自社の経験などと照らして妥当かなどといった判断を通して、モデルの妥当性・信頼性の向上を図っていく姿が求められるかと感じております。次のページ以降では、得られたDAGを用いて、CPTの推定のあとに、事後確率および因果効果の計算をして、入院発生率にどの程度影響があるかを評価しました。



まず、こちらのページで、CPTの学習結果についてまとめております。ただし、こちらは、本日は最尤推定により学習した結果のため、データセットの割合がそのまま学習された結果となっております。本日は、時間の関係で、こちらの最尤推定の結果については省略させていただければと思います。

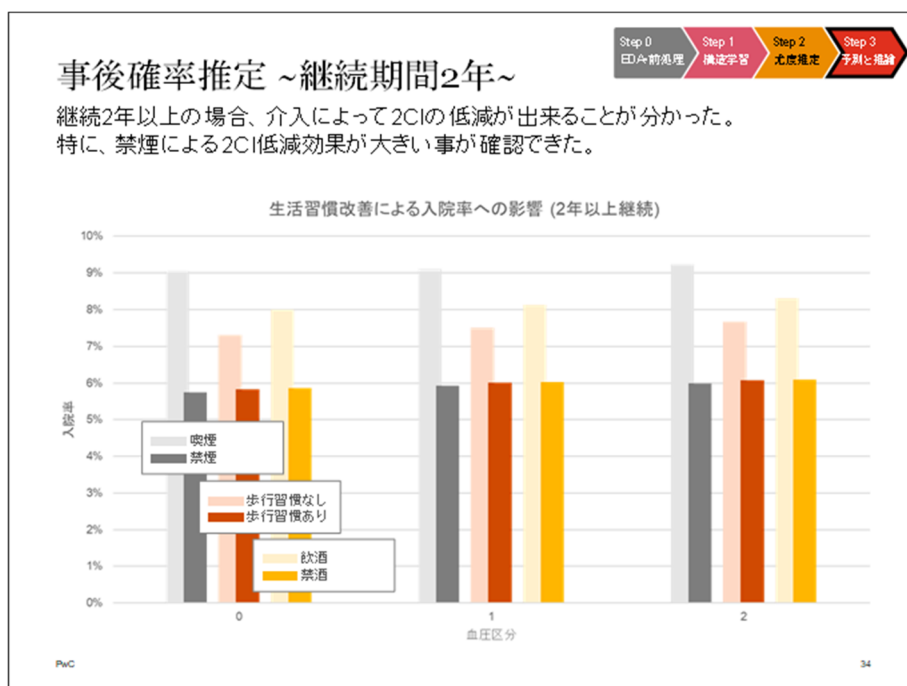


次に、CPTの学習が完了しますと、事後確率や因果効果の推論ができるようになります。例として、飲酒習慣で条件づけた場合の入院の事後確率を紹介いたします。

一番左の列には、単なる事前分布として各ノードの周辺分布を記載しております。そして、1番の飲酒習慣のノードに飲酒という条件づけを行った場合の事後確率が、こちらの列に記載されています。この「エビデンス」と書いているところは、確率変数に条件づけをするといったような意味で使用しております。飲酒習慣のノードを飲酒で条件づけた場合に、子ノードである喫煙習慣や、その他の健診項目などの確率をアップデートして、最終的に入院発生率の確率をアップデートするといったようなイメージをしていただければ

ばと思います。

反対に、禁酒で条件づけた場合の入院発生率がこちらに得られております。こちらの例では、飲酒している集団と禁酒している集団の入院発生率には、2%ほどの入院率の差があるような結果となりました。以降のページでは、このような事後確率と因果効果の二つの推定結果を比較しております。前半でお話したように、こちらのページのような単純な事後確率には、交絡因子による影響が含まれている可能性がございます。実際、今回もそのような影響が見られており、事後確率と因果効果を見比べることで、事後確率による施策の評価を行うことの危険性もお伝えできるかと思います。



まず、事後確率の推定結果から紹介いたします。こちらのページに記載している結果は、継続期間2年のものになります。本日は、この横軸にあるように、血圧で条件づけした場合の、各生活習慣の改善がどの程度あるかを評価した結果を紹介しております。血圧ごとに評価した背景としまして、各被保険者の体況に応じて、介入による効果の差が見られた場合に、個人ごとに最も効果的な生活習慣の改善を提案するといった応用もできると考えて、このような分析を行っています。このグラフでは、薄い棒グラフが改善前の入院率となっておりまして、濃い棒グラフが改善後の入院率となっております。こちらに示しているとおり、介入によって入院発生率が減少しているということが確認できます。特にグレーで示している喫煙について、大きいという結果となっております。これは、構造学習の結果で見たとおり、禁煙の習慣から入院へのエッジが継続2年以降で学習されたことによって、禁煙の習慣が入院率へ直接影響を与えているためと考えられます。

事後確率推定

事後確率および生活習慣改善前に対する改善後の入院率の割合をまとめた。
継続1年では各生活習慣の改善が入院率へ与える影響はほとんどなかった。
一方、長期間継続することで、入院率へ与える影響が大きくなることが観察された。

継続	血圧域	喫煙	禁煙	低減割合	歩行習慣なし	歩行習慣あり	低減割合	飲酒	禁酒	低減割合
1年	正常	7.0%	7.0%	0.0%	7.0%	6.9%	-1.5%	6.9%	7.0%	1.6%
	高値正常	7.2%	7.2%	1.0%	7.3%	7.2%	-2.0%	7.2%	7.3%	1.6%
	高血圧	7.3%	7.3%	0.3%	7.4%	7.2%	-2.1%	7.2%	7.4%	2.0%
2年	正常	9.1%	5.7%	-36.6%	7.3%	5.8%	-20.1%	8.0%	5.9%	-26.7%
	高値正常	9.1%	5.9%	-34.9%	7.5%	6.0%	-19.9%	8.1%	6.0%	-26.0%
	高血圧	9.2%	6.0%	-35.1%	7.7%	6.1%	-20.7%	8.3%	6.1%	-26.9%
3年	正常	8.1%	5.5%	-32.6%	7.2%	5.5%	-24.0%	7.6%	5.6%	-27.4%
	高値正常	8.3%	5.7%	-31.3%	7.5%	5.7%	-24.2%	7.9%	5.8%	-26.8%
	高血圧	8.4%	5.8%	-31.1%	7.6%	5.7%	-24.1%	8.0%	5.9%	-26.6%

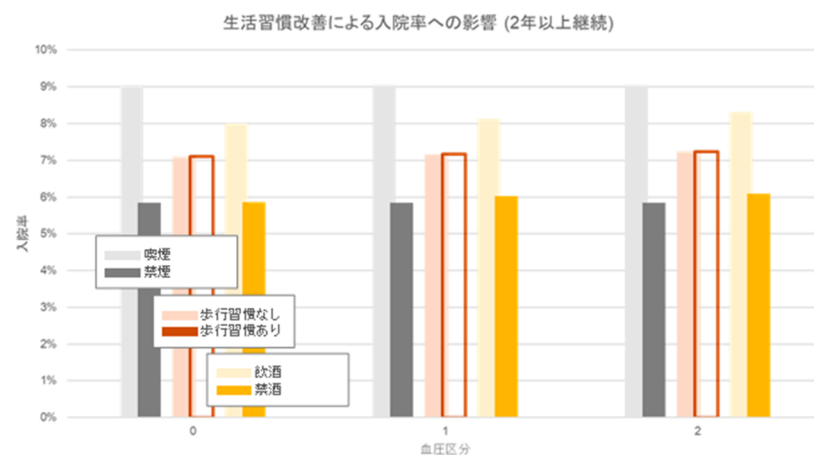
PwC

35

こちらは、継続1年から継続3年までの事後確率推定結果をまとめたものになっております。「低減割合」と書いてあるカラムについては、生活習慣改善前の入院率に対する、改善によって低下した入院率の割合を示しています。結果としましては、継続1年においては、どの生活習慣もほとんど効果は見られておりませんが、より長期間継続することによって、入院発生率へ影響を及ぼしているといったような結果となっております。

因果効果推定 ~継続期間2年~

歩行習慣から他へ伸びるエッジがなく、歩行習慣への介入による因果効果が0と評価された。



PwC

36

続いて、因果効果の推定結果を紹介いたします。特徴的な点として、歩行習慣の因果効果がゼロとなっている点が挙げられます。このような結果となった理由としましては、継続2年のDAGでは、歩行習慣からはどこにもエッジが伸びていなかったことが原因となります。因果効果を推定する際は、親ノードからのエッジを切断した確率分布を考えるということを説明しましたが、親ノードからのエッジを切断した結果、歩行習慣のノードは他のノードから独立してしまっ、エビデンスを与えたとしても他のノードへの影響はゼ

ロとなってしまったため、このような結果となっています。

Step 0
EDA前処理
Step 1
機械学習
Step 2
尤度推定
Step 3
予測と検証

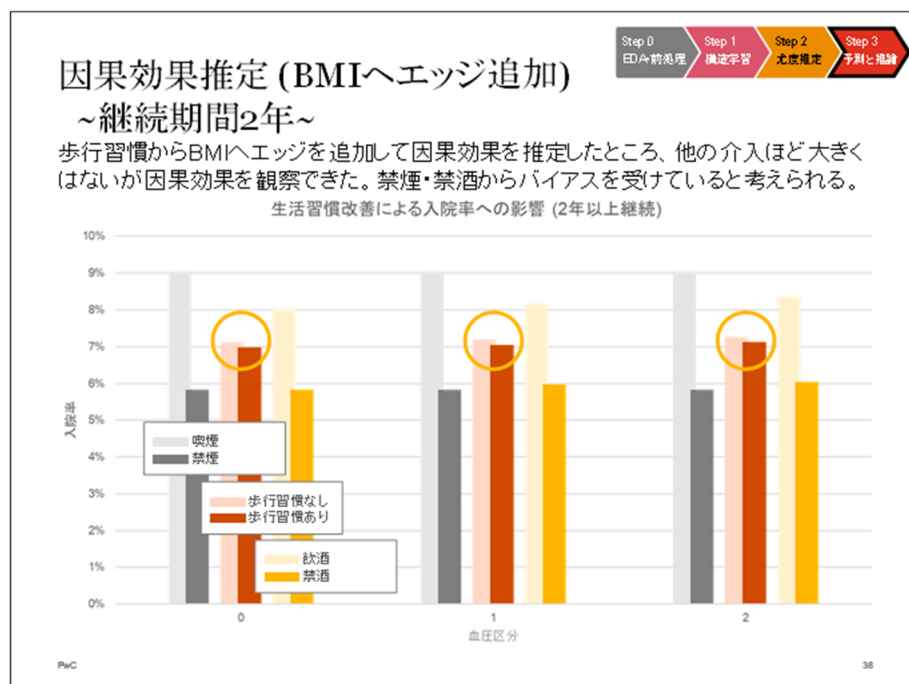
因果効果推定

継続1年から3年をまとめると以下の通り。
 継続2年では歩行習慣の因果効果が推定できなかった。継続1年、3年では歩行習慣からBMIへのエッジが学習された事から、継続2年でもエッジを引いて因果効果を推定した。

継続	血圧域	喫煙	禁煙	低減割合	歩行習慣なし	歩行習慣あり	低減割合	飲酒	禁酒	低減割合
1年	正常	7.1%	7.2%	0.6%	7.2%	7.1%	-1.9%	6.9%	7.0%	1.6%
	高値正常	7.1%	7.2%	0.5%	7.2%	7.1%	-1.9%	7.2%	7.3%	1.6%
	高血圧	7.1%	7.2%	0.5%	7.2%	7.1%	-1.9%	7.2%	7.4%	2.0%
2年	正常	9.0%	5.9%	-35.4%	7.1%	7.1%	0.0%	8.0%	5.9%	-26.7%
	高値正常	9.0%	5.9%	-35.4%	7.2%	7.2%	0.0%	8.1%	6.0%	-26.0%
	高血圧	9.1%	5.9%	-35.5%	7.2%	7.2%	0.0%	8.3%	6.1%	-26.9%
3年	正常	8.2%	5.6%	-31.7%	7.1%	7.0%	-2.3%	7.6%	5.5%	-27.4%
	高値正常	8.2%	5.6%	-31.7%	7.2%	7.0%	-2.3%	7.9%	5.8%	-26.8%
	高血圧	8.2%	5.6%	-31.7%	7.2%	7.0%	-2.3%	8.0%	5.9%	-26.6%

PwC
37

因果効果も、継続1年から継続3年まで比較した結果をこちらに掲載しております。継続1年においては、事後確率と同様にほとんど効果は見られませんでした。また、継続2年においては、ゼロとなっていた歩行習慣による因果効果は、他の継続期間において僅かに見られております。ただし、親ノードの飲酒習慣と喫煙習慣が交絡因子となっておりまして、因果効果は小さいものとなっております。実際には、継続期間2年においても、歩行習慣から入院への因果効果はゼロでない想定されるため、継続1年および3年のDAGで学習されていた歩行習慣から、BMI、エッジを継続2年のDAGでも追加して、因果効果を推定した結果を確認していますので、そちらの結果を紹介いたします。

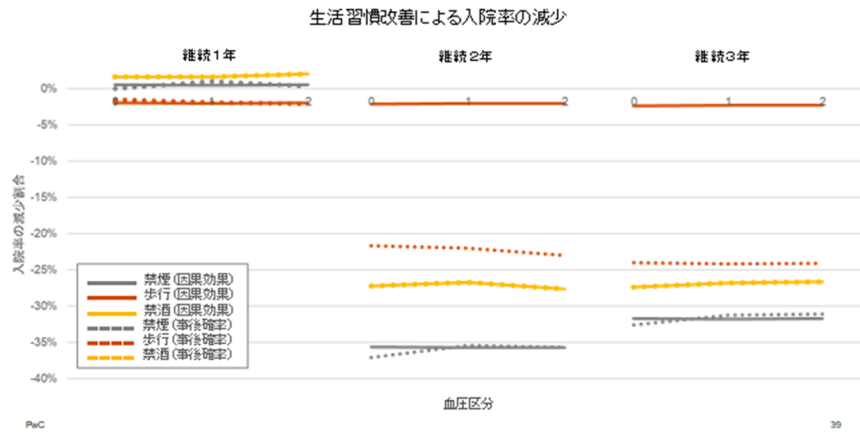


こちらが、更新後の因果効果の推定の結果になります。新たに歩行習慣からBMIにエッジを引いたことで、歩行習慣による因果効果が僅かに観察されるといった結果となりました。

因果効果 vs 事後確率

Step 0 ED/A 前処理 Step 1 横断学習 Step 2 尤度推定 Step 3 予測と推論

- 事後確率と比較すると、歩行習慣は低減割合が小さくっており、交絡因子の影響を受けていることが分かった。
- 禁煙、禁酒は2年以上継続した場合に入院率が大きく低下するため、介入施策として有用であると考えられる。



因果効果と事後確率を比較した結果を最後に紹介いたします。このグラフでは実線が因果効果を表しておりまして、点線が事後確率を表しております。このように比較すると、歩行習慣は、事後確率に比べて因果効果はかなり低く評価されているということが分かります。これは、歩行習慣自体は入院率へ与える影響は少なく、その交絡因子である喫煙習慣や禁酒習慣による影響が大きいのではないかと考えられています。

こういった結果を見ると、事後確率で施策の評価を行うことは危険であり、施策の比較を行うためには、因果効果を用いた比較を行うことが適切であると考えられます。ただし、この歩行習慣については、今回は自己申告による問診項目の回答のため、ウェアラブル端末などの歩行のデータを用いることで、もう少し因果効果が見られるといった結果も考えられますので、そちらを今後検討させていただきたいと思っております。また、体況ごとに施策の効果が異なるのかといった問いも立てていましたけれども、横軸の血压区分ごとの差についてはあまり見られないといった結果となりました。

因果推論のまとめ

Step 0 ED/A 前処理 Step 1 横断学習 Step 2 尤度推定 Step 3 予測と推論

結果

- DAG構造からわかるバイアスの影響を調整した因果効果の計算により、施策の比較を行い、長期間の禁酒・禁煙の効果による効果がみられた。
- 歩行習慣は禁酒・禁煙が交絡因子となっており、事後確率と比較すると因果効果は小さく評価された。

留意点

- 歩行習慣の例のように、DAGの構造によっては因果効果が0になり正しく計算できないモデルが得られる事もある。
- このような場合、アクチュアルな判断を基にDAGを見直し、より適切に因果効果を評価することが重要である。

PwC

40

このページで、因果効果の推定の結果と留意点をまとめます。結果については先ほど紹介いたしましたので省略させていただいて、留意点の方を説明させていただきます。まず、歩行習慣の因果効果がゼロとなってしまうように、因果効果はDAG構造により大きく影響を受けることが確認できました。このような場合、因果効果は本当にゼロなのかどうか、ゼロでない場合にはどのようにしてDAGを見直して、因果効果を推定するのかといった判断を行うことが必要となります。DAGの見直しについては、本日は他の継続期間のDAGを参考にして、継続期間2年を見直しましたが、実際にはアクチュアリーや社医等の専門家による意見を踏まえて、構造学習にて与えられた事前知識が適切かなど、プロセス全体を通して判断を行うことが必要と考えております。

医療データの適用例の説明は以上となります。最後に全体のまとめをいたします。板倉さん、お願いいたします。



板倉 はい、ありがとうございます。それでは、本日の内容を振り返りたいと思います。

本日のまとめ

- 保険前提を介入できる変数とみなす事で、利用可能な医療データの増加に伴ない、保険会社で活用できるシーンはさらに増える。
- 因果探索手法を導入する事で、リスク因子間の因果関係を考慮して個別保険リスク評価を精緻化し、リアルワールド評価を行うことができる。
- この手法により介入効果を評価する事で、従来の保険数理手法に残る交絡因子の効果や加算評価の問題などのミスプライス要因を排除できる。
- 実務上は、真のDAG構成に向けた専門家知見の活用、モデル結果の安定性、被保険者群団との差異を踏まえた効果測定など様々な課題がある。
- 特に安定的な結果を得るには、複数年での結果比較やアウトカム期間の検討など、経験値分析実務を踏まえたアクチュアリーによる判断が重要になる。
- 分析結果を踏まえて意思決定を行う際には、得られたDAG構造や因果効果の解釈を踏まえてステークホルダー間で調整する。事前知識と反していないか、反している場合はその理由を把握する検証姿勢が求められる。

PwC

42

まず、保険前提を介入できる変数と見なしたというところで、今、特に、先ほど少しお伝えした、利用可能な医療データが増えていると。高齢の群団のところで体況が見られる、死亡の転帰が使えるようになるといったようなところで、保険会社さんで医療データを活用するシーンがさらに増えると思っております。さらに、今日ご紹介したような因果探索手法を導入することで、リスクファクター下の因果関係を考慮した、より個別的な保険リスク評価で精緻化して、リアルワールドでの評価に近づけることができます。

この手法を用いるメリットなのですが、これによって、従来、保険数理手法で残ってしまうような交絡因子の効果を見逃してしまったり、あと、先ほどの加算評価の保守性というような問題がミスプライスの要因にもなりうるのですが、こういったところが排除できるというのがよいところだろうと思っております。

一方で、実務上なのですが、真のDAGがもし仮にあるとしたら、これを構築するに向けて、いろいろな専門家と、先ほど社医を代表として挙げましたが、そういった人と知見を活用したり、モデルの結果を安定させるように制約を加えたり、被保険者群団、これは今、医療データを用いたプールを仮想的に作ってやったということなのですが、それを自社の被保険者群団との差異がどれくらいあるのかなどといったもので効果測定をしていくところに、さまざまな課題が出てきやすい、とても難しい手法だろうとは思っています。

一方で、アクチュアリーがそこに積極的に関与していくというところで、例えば、経験値分析手法の実務から類推されるような手法、結果として、そこにアクチュアリーが中心的に関与していくということが非常に重要なのだらうと思っています。先ほどの歩行習慣が1年、3年で因果関係があるのだけれども、2年で出ないなどといったようなときに、では、DAGの因果関係を、ここを引くのだというような判断はアクチュアリーにしかできないのではないかと思いますし、実際にその結果を安定させるというところは、こういったジャッジメントが非常に重要になります。

また、その結果を踏まえて意思決定をするといったときも、では、DAGの構造に、どこがどれだけ依存しているのか、因果効果はどうやって解釈するかというところを、またステークホルダー間、意思決定者とステークホルダーと、それぞれお話をきちんとしないといけないので、そういった今まで知られている知識

と反しているのか、反していないのか、反していれば、どこが違う要素なのかというようなことを、検証したうえでお伝えしていくということが非常に重要です。一方で、今まで得られていないインサイトもきちんと得られるので、保険会社さんでも広く使われてほしいと思っているところでございます。

では、本日は、われわれの発表としては以上になります。ご清聴いただき、ありがとうございました。

滝瀬 では、質疑応答の時間に入らせていただきます。スライドでもかなりの質問をいただいていますけれども、まず、会場の方から質問を受けさせていただきます。質問のある方は、挙手をお願いいたします。

カネムラ カネムラといいます。本日は興味深い発表をありがとうございます。

説明の後半でずっと強調されていた、介入で禁酒、禁煙、歩行習慣、あるいは歩行習慣から伸びていないという話をされていたのですが、よく分からなかったことは、例えば32ページにあるように、歩行習慣からエッジが伸びていないのは事前知識の与え方（注：当日は「DAGの設定」と質問頂いた）次第で、エッジが伸びることになるのでしょうか。それとも、そうではなく、事前知識の与え方いかんにかかわらず、歩行からエッジが伸びないのでしょうか。

板倉 ご説明する中で、若干説明不足だったかもしれないので、再度お伝えします。事前知識として、矢印の向きがアウトカムから健診結果や生活習慣などに行かないといった制約は原則加えてはいるのですが、観察データからモデルで学習させて、出てくる因果関係を大事にしています。要は、データから出てくるインサイトが何かというところで矢印を引いているというものがそれに当たりますので、恣意的にどこからどこに引くというようなことは、原則やりません。

カネムラ おっしゃっていることとしては理解しているのですが、事前知識の与え方によっていろいろと、その前にやった、スコアは幾つだからこの線になるとか、事前知識の与え方によって、結果、そのスコアの部分が違って、矢印がどう出ることが変わるかと思うので。だから、GPTやGOTなど、そういったさまざまな事前知識の与え方の設定のやり方によっては、結果的に歩行習慣から伸びるということもあるのか。たまたまこの事前知識の与え方によって、歩行習慣が伸びなかったのかどうかということなのですね。

それとも、あそこのDAGのオレンジ色の設定はどのようなことをやっても、大体微妙な差は出るかもしれないけれども、歩行習慣が伸びることが結構ありうることになるのでしょうか。

板倉 ひょっとして設定とおっしゃっていることは、多分、離散化や各変数の設定の仕方によっては、因果関係のDAGが変わるのではないかということでしょうか。

カネムラ 極端に言えば、HDLから始まる10個ぐらいのエッジをなくして、いきなり上の三つから一番下の2CIまでの部分に設定したら、歩行習慣から伸びるというようなことがありうるかどうか。

堤 私の方から回答させていただきます。これ自体はやってみないと分からないというところがございまして、事前知識に与えたエッジや、あと、離散化の方法などによって少し結果が変わってきて、結果的に歩行習慣からBMIへエッジが伸びるといった結果も得られることは考えられます。

カネムラ ということですよね。もう一つ続いて技術的な話。そうすると、でも、オレンジ色の設定によっても、多少かもしれないけれども、多少変わるぐらいで、あまり大きく、矢印がぐっと、太い線と言え少し変ですけども、いきなり上がるというようなことは多分なく、オレンジのグラフの設定によっても、影響は、微細にずれるかもしれないぐらいの分だと理解していいのですか。

堤 そうですね。周りの状況に応じて変化すると思っていただければ大丈夫です。

カネムラ ありがとうございます。

板倉 ありがとうございます。モデルがロバストではないという点で言い切れない部分もありますが、少し変えたぐらいでは、それほど矢印の因果関係が大幅に変わるということはないですね。

カネムラ ありがとうございます。

滝瀬 では、スライドで幾つか質問をいただいていますので、こちらを取り上げさせていただきたいと思います。まず、「健診時点で70歳の女性を対象に分析したとありますが、年齢、性別もノードとして考えて分析することもできるかと思います。ノードに入れなかった理由は何かありますでしょうか」という質問です。

堤 私の方から回答させていただきます。

おっしゃるとおり、こういった分析もできるかと私たちも考えたのですが、1点は、まず、結果の解釈が少し難しくなってしまうかなということと、あと、年齢ごとにデータの中に含まれるボリュームゾーンが結構変わってしまうということもあって、そういったボリュームゾーンに結果が引っ張られることも考えられましたので、今回は分かりやすさのために、いわゆる層別解析のように行っております。ただ、そういったことも考えられると思いますので、今後、検討していきたいとは考えております。

滝瀬 はい、ありがとうございます。

では、もう一つ。「DAG構造を前提とした測定ですが、結果導出後、その妥当性検証をどのようにするのでしょうか。例えばDAG構造を前提としていたが、実は非DAG構造で因果がループしていたことなど、検出できるのでしょうか」という質問です。堤さんへの質問ということです。

堤 ありがとうございます。ループについて少し答えさせていただければと思います。そもそもループの検出は、現在の手法では検出されません。ループがあるものについては、事前にそのデータをよく観察して、ループがあることを判断したうえで、ループを扱うようなモデルを使用することの方が適切だと考えております。ループがあるような場合は別のモデルもございまして、現在のわれわれの環境ではそういったモデルは考慮できておりません。無向グラフなどを使用して、巡回があるものを表現できるものもあるようなのですけれども、現在は考慮できていないため、そういったところも考えて、ループを処理するにはどうするのか、今後の検討内容かと考えております。

滝瀬 はい、ありがとうございます。すみません、まだ多数いただいているのですけれども、もうお時間が来てしまいましたので、以上をもちましてセッションC－1は終了させていただきたいと思います。

発表者の板倉さん、堤さんに、いま一度大きな拍手をお願いいたします。

板倉 ありがとうございました。