原著論文

乳がん・前立腺がん経験者のインタビューテキストデータから集団機械学習ランダムフォレストによる検診行動の推定の試み

——DIPEx-Japan のテキストデータ二次分析——

木 村 朗[1)]

Predicting screening actions by quadrat analysis using artificial intelligence, and from the text data of cancer screening based on the interviews of patients who had an experience of having a breast cancer and a prostate cancer in Japan

Akira KIMURA[1)]

要 旨

本研究の目的は、厚生労働科学研究補助金がん臨床研究事業の一環として作成され、DIPEx-Japan が管理する、わが国の乳がん・前立腺がん経験者の語りのインタビューを基にがん検診に関するテキストデータから人工知能を利用した二次分析によって検診行動の推定の可能性を明らかにすることであった。対象は疾患特性・地域特性等を考慮した、対象者の多様性を確保するサンプリング法（Maximum variation sampling survey）で得られた乳がん経験者38例、前立腺がん経験者46例のデータであった。方法はテキストデータに対して人工知能を利用する集団機械学習ランダムフォレスト法を用いて gini 係数を基に作成したモデルから推測した固有名詞を用いた予測成績を求めた。結果として、gini 係数により検診受診の有無の鑑別成績を高めたものは、前立腺がん経験者で「サプリメント」、「PSA」、乳がん経験者で「マンモ」、「浮腫」という順であった。モデルの推測成績は前立腺がん経験者で47.6%、乳がん経験者で59.5%の判別性を示した。人工知能による集団学習と機械学習によって生成したモデルは、両者の間で医学専門用語と一般用語の頻度の比において逆転していた。二次分析手法に人工知能を用いることで、このような知見を得られる可能性があることから、データマイニングをインタビュー開始から間もない時期に行うことで、より適切な情報から検診行動を推定するための構造化質問の作成が容易になる可能性が示唆された。

キーワード：検診行動推定，ランダムフォレスト機械学習，テキストマイニング，乳がん経験者，前立腺がん経験者

I 研 究 背 景

「癌」に関する専門家向け診療情報・医療情報に較べ、患者自身の経験や生活機能に照らして患者の QOL を向上させる上で欠かせない情報の不足は国際的にも、国内的にも課題となっている。また、それらの情報の正しさや利益相反に照らしたプロセスで開発され発信されることが望まれている。今日、Evidence

1）群馬パース大学保健科学部理学療法学科

Based Medicine の実践は、国際的な標準になっており、それを補塡する Narrative Based Medicine のために患者自身の病いの経験を集めたデータベースが作られている。その一つで、国際的に展開されている Oxford 大学と NPO が開発した Database of Individual Patient Experiences（個々の患者の体験のデータベース）DIPEx がある。

日本では、和田らによって DIPEx の手法を用いた、患者の病いの経験を動画およびテキストでデータベースの作成が、厚生労働科学研究助成を受け開始された[1)]。

さらに、患者のがん情報の不足を補うことを目的とした、これらのデータシェアリングに関する研究が中山らにより2010年度より厚生労働科学研究補助金がん臨床研究事業の一環として開始された。これらのデータは日本において組織された DIPEx-Japan によって管理運営されている[2-4)]。さらに、厚生労働科学研究班の朝倉隆司らの下、我々はこれらのデータベースの活用のための二次分析方法として、テキストデータから単語の頻度や品詞、感情を表わす形容詞の頻度から検診行動に関する分析を試みた。

しかし、単語の基本統計量の集計からは、検診行動を推定しうる結果は得られなかった。

そこで、人工知能を利用したデータマイニングを試みた。ここでデータマイニングとは既知のデータからモデルを作成し定義された問題の答えを導き出すことと定義する。未知のことを予測する際に、知りたいことが分類を通して得られることか、回帰を要することの両者のうち、どちらかを用いる。

2013年より、手法の根本的な見直しを行い、有償ソフトウェアでは、それを持つ人以外にデータマイニングの検証が難しいという点を踏まえ、誰でもいつでも可能な方法を取り入れることで、より客観性を高めることを目指した。

作成するモデルとして、まず、樹木モデルの適応を考えた。いわゆる人工知能による自動分類判断の操作を行うための樹木モデルでは、菖蒲のデータを品種ごとに分類する決定木の例が説明に用いられる。葉の長さや葉の幅の違いだけから品種を推定するというものである。

このプログラムは1960年ごろに開発され、C4.5と呼ばれるモデルを1986年 Ross が開発[5)]、CART のアルゴリズムを Breiman らカリフォルニア大[6)]、Freidman らスタンフォード大の研究者によって公開された[7)]。本研究では、CART のアルゴリズムを弱学習器（少ない変数からなる多数の回帰式を作成し、求める推定の成績を高めたモデルに貢献した変数の効果を調べるもの）として使用し、分岐ルールに gini 係数を用いる集団学習を行うアルゴリズムを用いたモデル生成を試みた。樹木モデルは、Tree-based model 非線形回帰分析、非線形判別分析の１つの方法であり、説明変数の値を何らかの基準をもとに分岐させ、判別・予測のモデルを構築するものである。分岐の過程は木構造で図示することができる。分岐ルールは分類器とも表現される。この分岐ルールに gini 係数（図１）を用いることができる。さらに、モデルの特徴として、IF-THEN のようなルールで表すことができる。これらは、理解しやすいため、最もデータマイニングの中で応用されている。

$$\text{entropy} = -\Sigma p\,(i \mid t)\,\log^2 p\,(i \mid t)\ (i = 1\text{to } c)$$
$$GI = 1 - \Sigma\,[p\,(i \mid t)]\ 2\ (i = 1\text{to } c)$$

図１　gini 係数を求めるための樹木モデルのアルゴリズム
この樹木モデルを500回のバギングによって最も目的行動の分岐を高い確率で示す gini 係数を持つテキスト（語）の発見を集団学習・機械学習を通して行う[8)]。がん経験者のインタビューテキストデータから得た集団機械学習の結果を表わしている。

II　目　　的

本研究の目的は、DIPEx-Japan のデータベースと実際にインタビューを行ったインタビューアーの持つデータを合わせた二次分析用データから、統計言語である R のバギングシリーズを利用し、コンピュータによる人工知能を使った集団機械学習から検診行動の有無を推定するモデル作成を試み、このモデルの成績を明らかにすることである。

III　対　　象

分析対象は DIPEX-J（前立腺がん・乳がん患者の語り、以下 PC，BC とする）[9)] の二次データである「がん経験者の語りのテキストデータ」と実際にインタビューを行った「インタビューアーの持つデータ」であった。これらの「がん経験者の語りのデータ」は和田らによりテレビ、新聞、HP、マスメディアおよび

ヒューマンリレーションによって疾患特性・地域特性等を考慮した、対象者の多様性を確保するサンプリング法（Maximum variation sampling survey）で得られた PC38例、BC46例であった。

分析対象例数は、検診受診行動の有無に関わらずテキストデータが存在する PC38例、BC46例であり、インタビューアーによる検診受診行動の有無の情報を追加した上で、全例のテキストデータを解析に使用した。

倫理的配慮について、これらのデータの使用、解析、公開にあたって DIPEx-Japan とデータベースの借用契約を結び、その際に DIPEx-Japan の倫理委員会による審査が行われ、研究実施の承認を得た。

これらの一次データは実際の運用開始に先立ち、インタビューデータをすべて匿名化し、本人によるチェックで公開を希望しない部分の編集削除が行われた。さらにインタビュー協力者の個人情報保護と、インタビュー協力者と研究班の両方に帰属する著作権の保護に配慮した「データシェアリング規定」が作成され、シェアリング希望者から提出された申請書（研究計画書）を、「がん患者の語りデータベース」研究班の委任を受けた「情報倫理委員会」が審査した上で、データの貸出が行なわれた。

Ⅳ 研 究 方 法

2010年から2012年にかけて、DIPEx-Japan によって収集された患者の語りに関する動画および音声データより、半構造インタビューに関するテキストデータ（以下、二次データ）を、スタンドアロン型コンピュータに取り込み、奈良先端科学技術大学の開発による chasen 2 を用いて形態素要素に分解した。同時に、インタビューアーから追加情報を得て、ケースごとに属性情報を対にした（データクリーニング後データ、以下後データ）。このデータに対し、表計算ソフトによる関数式を用い、ipadic2.0（奈良先端科学技術大学）を利用して特定の単語の頻出量を求めた。

特定の単語を説明変数として、頻出量を基に、統計言語Ｒを用いて樹木モデルを作成した。起点となる変数について中央値を用いた分類を試み、gini 係数の高い語を求めた。

DIPEx-Japan が管理するテキストデータからの解析用データセット作成手順

1．WinCha 2000および Chasen 2（奈良先端科学技術大学）[10]、形態素解析にてケースごとの頻出単語を抽出し、品詞分類の数量統計および ipadic2.0（奈良先端科学技術大学）[11] の辞書にない頻出固有名詞上位３語（以下、分析語）を求める。

2．randomForest 法（以下 RF、R パッケージ ver. 3.0.1、OS は linax.ubuntou）* による分析語の量的分布の差異を利用した判別（検診受診の有無）による RF 決定木モデルを作成する。反応変数として受診行動の有無をインタビューアーより取得し、機械学習の際に教師付き条件でモデルを生成する。

3．RF 決定木モデルの判別確率を計算する。

4．RF 決定木モデルにおける分岐ルール（gini 係数ー投票数由来）における最も有効な分析語を発見する。

＊ randomForest は2001年に Breiman 氏が提案した新しいデータ解析の方法である。

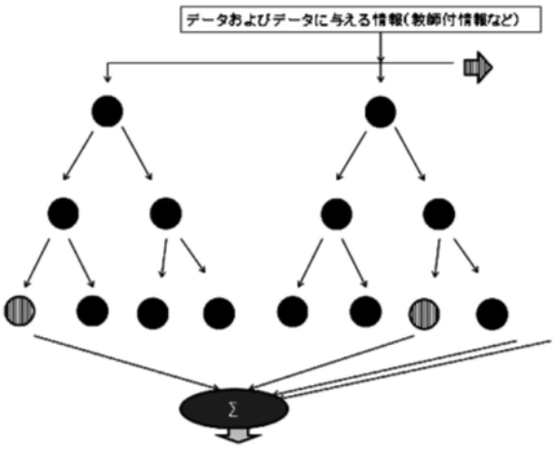

図２ randomForest による集団学習・機械学習の概要
図は、少ない変数からなる多数の回帰式を作成し、求める推定の成績を高めるモデルに貢献した変数の効果を多数決で決定する randomForest のイメージをあらわしている。分岐ルールに gini 係数を用いる集団学習を行うアルゴリズムを用いたモデル生成を行う樹木モデルは、Tree-based model 非線形回帰分析、非線形判別分析の１つの方法であり、説明変数の値を何らかの基準をもとに分岐させ、判別・予測のモデルを構築する。

Ⅴ 結　　果

商業マイニングソフトウェアを用いない無償ソフトウェアＲで開発配布されている randomForest を使用する本研究で示した操作によって、患者の病いの語りデータベースの二次分析としてのテキストマイニン

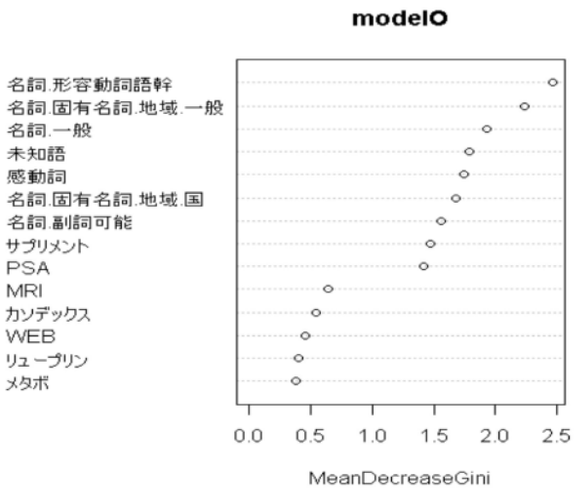

図3　前立腺がん経験者の検診受診行動推定に貢献するテキストの gini 係数
前立腺がん経験者のデータセットから 2 組のサンプルを作成し、4 個の変数をサンプリングした。これらから決定木をつくる過程をおよそ500回繰り返して得られた。この過程で量産された決定木のすべてに対して、予測したデータを入れ、この結果の多数決をとり、予測結果とした中で、有効な変数の gini 係数を横軸に、その数値を示した変数（固有名詞）を縦軸にプロットしている。

グ手法は、計算結果を出力することに成功した。

PC と BC の検診受診予測モデルは PC が図 3 と BC が図 4 に示すようになった。

これらの図は、データセットから 2 組のサンプルを作成し、4 個の変数をサンプリングした。これらから決定木をつくる過程をおよそ500回繰り返して得られた。この過程で量産された決定木のすべてに対して、予測したデータを入れ、この結果の多数決をとり、予測結果とした中で、有効な変数の gini 係数を横軸に、その数値を示した変数（固有名詞）を縦軸にプロットしている。

分岐ルールの弱学習器の集合体から得られた gini 係数に基づく、PC、BC のがん経験者において、検診行動の実行性を高めたものは、ipadic2.0で非固有名詞となる単語として、PC では、「サプリメント」が最も大きく、次いで「PSA」、「MRI」の順であった。一方、BC では、「マンモ」が最も大きく、次いで「浮腫」、「ブログ」という順であった。

PC モデルの判別性能は、以下の様に出力（R の出力結果のまま）された。

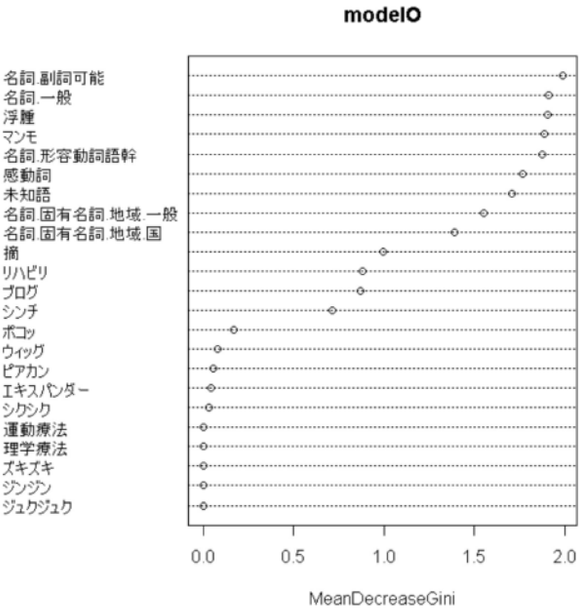

図4　乳がん経験者の　検診受診行動推定に貢献するテキストの gini 係数
乳がん経験者のデータセットから 2 組のサンプルを作成し、4 個の変数をサンプリングした。これらから決定木をつくる過程をおよそ500回繰り返して得られた。この過程で量産された決定木のすべてに対して、予測したデータを入れ、この結果の多数決をとり、予測結果とした中で、有効な変数の gini 係数を横軸に、その数値を示した変数（固有名詞）を縦軸にプロットしている。

誤差の推定値

OOB estimate of error rate : 52.38%
（筆者加筆、誤り率の推定値＞正解率47.6%）

Confusion matrix :

	n	y	class.error
n	12	11	0.4782609
y	11	8	0.5789474

BC モデルの判別性能は、以下の様に出力された。

OOB estimate of error rate : 40.48%
（筆者加筆、誤り率の推定値＞正解率59.5%）

Confusion matrix :

	n	y	class.error
n	15	8	0.3478261
y	9	10	0.4736842

Ⅵ　考　　察

モデルは、PC において、「サプリメント」、「PSA」という固有名詞が検診受診歴の有無の分類器として有

意な gini 係数を示した。同様に、BC において「マンモ」、「浮腫」という固有名詞が検診受診歴の有無の分類器として有意な gini 係数を示した。ランダムフォレストによる弱学習器による解析はこれらの語句の存在を示した。これらの語句は、医療従事者が構造化インタビューもしくは半構造化インタビューを行う際に役立つ可能性がある。具体的には、これらの語句から想起される概念は、患者の病の経験者としての生活上の困難を ICF などに従った個人因子、環境因子を特定する目的で質問文の作成、設問設定へのヒントになる可能性がある。人工知能型テキストマイニングによるモデルは、その時点で特定されていない固有名詞そのものや、固有名詞の組み合わせから導かれる概念の抽出において、初学者にヒントを与えるものと思われる。

また、これらの語句は、質的なデータを解析する場合に、従来のグランデッドセオリーなどの経験者の主観的な意味づけやカテゴライズ手法において習熟した指導者が得られない場合に有効であろう。初学者がミーニングに基づき、語りデータにおいてパラグラフの分類を考えたプロセスの説明を求められる際に、gini 係数の高い固有名詞を ipadic2.0（奈良先端科学技術大学）の辞書を基準に用いてキーワードを絞り込むことができる。この工程における可視化的な共有が可能になり、この作業における時間の短縮や、初学者の学習に貢献することが期待できる。

例を挙げれば、本研究のモデルから得られた知見は推定性能の評価が可能である。この知見で得られた PC と BC の経験者の語りの中の語の属性は、検診行動の分岐を決定する gini 係数が高い順に PC では「サプリメント」＞「PSA」となっており、これは一般用語＞医学専門用語であるのに対し、BC では、「マンモ」＞「浮腫」＞「ブログ」と、医学専門用語＞一般用語の順になっている。すなわち、それぞれのがん経験者の語りの特徴が可視化されている。検診行動を推測するには、がん経験者では語りの中の患者の発声する語句中の医学専門用語と、一般用語の区別に注意を払うことで、検診行動の有無を意識したインタビューを展開することが出来る可能性を示唆している。

このように、がん経験者の語りをテキスト化した二次的データは形態素要素に分解されることにより、集団学習・機械学習アルゴリズムを使った樹木モデル、ランダムフォレストによるモデル作成に用いることが出来ることが示された。

最後に、実用性の観点から重要なことはこの情報生成にかかるコストと時間である。無償ソフトでありながら、統計言語Rを用いたランダムフォレストの利点は、多くのデータセットを用いることによって、正確な分類を行うことができる。このようにデータマイニングにおける分類問題において、説明変数の重要度を見積もることで、従来のテキストマイニングに比べ時間コストを大幅に減らしている。また、欠損したデータを良い精度で推測できるので、データの大部分が欠損していても正確さを保つことができるとされ、従来の手法に比べても、学習速度が早いことから、処理速度の速さで知られる google におけるスパムメールの判定に用いられている位、性能が良い。これらの点で、集団機械学習ランダムフォレストの実用性は高いものと考えられる。

モデルの推測成績は PC で47.6%、BC で59.5%と BC の方が優れた判別性を示した。これは、PC の方が BC よりも例数が10上回ったものの、語りの単語数の絶対数が多いという、単純な推計統計上の有利な条件を抑え、治癒可能性に関して BC の方が厳しいことが情報として日常生活の中で、容易に得られる可能性が考えられる。また、医学専門用語と一般用語の発語頻度の比が、PC と逆になっていることから、古典的保健行動理論の視点で考えた場合、危機回避のために専門的知識を得ようとする、危機意識の高さによる受診行動の促しに性差がある可能性が推察され[12,13]、これらの理論に加え、性に関連する生活機能への影響が検診行動に関連している可能性も考慮した保健行動理論の形成に役立つ知見が得られた可能性がある[14]。

従来、カテゴリー分類を通して、意味づけを行う作業に代表される保健行動の質的研究を主とした患者の語り、経験談のインタビュー分析研究から、本研究が示す、人工知能を用いた集団学習・機械学習によるモデル生成による人間の認知機能の補助を果たしうる質-量的研究が可能になったことは、新たな保健学領域の研究方法のバリエーションを拡げたと考える。本研究で用いた方法は、今後の保健学、看護学、リハビリテーション科学等、臨床科学におけるエポックの１つとしても興味深い知見をもたらした。

Ⅶ 結　　語

前立腺がん、および乳がんの経験者の語りデータの二次分析として人工知能による集団学習と機械学習によって作成したモデルは、検診受診の有無の推定とし

て前立腺がん経験者で47.6%、乳がん経験者で59.5%の正解率を示した。　両経験者の間では医学専門用語と一般用語の頻度の比において逆転していることが明らかになった。このような知見を得るデータマイニングによって、より適切な情報から検診行動などの保健行動を推定するための半構造化質問の作成が容易になる可能性が示唆された。

追記

本研究で使用したDIPEx-Japanのデータ借用の契約書第12条「乙は成果物を公表する前に語りデータが適正に利用されていることを甲に示し、公表の許可を得る。」に従い、2014年3月に開催されたDIPEx-Japanの開催する委員会において、本研究の内容を発表し、論文化を進めることを確認していることを、ここに明記する。

謝辞

本研究は、中山健夫．平成22年度厚生労働科学研究費補助金第3次対がん総合戦略研究事業「国民のがん情報不足感の解消に向けた「患者視点情報」のデータベース構築とその活用・影響に関する研究の一環として質的データ分析に基づいた患者視点情報のデータベース化とデータシェアリングを通じた質的データの活用に関する研究の一環として、筆者が研究協力者として、NPO法人健康と病いの語りディペックス・ジャパンの二次データベースを利用して行った研究である。同研究代表中山健夫先生、研究班長朝倉隆司先生、佐藤（佐久間）りかさん、射場典子さん、澤田明子さん他、二次データの基となる語りを提供して下さったがん経験者の皆様に心より感謝申し上げます。

文　　献

1）和田恵美子．厚生労働科学研究補助金がん臨床研究事業「がん患者の意向による治療方法等の選択を可能とする支援体制整備を目的とした、がん体験をめぐる「患者の語り」のデータベース」平成21年度総括・分担研究報告書．2010年．
2）中山健夫．平成22年度厚生労働科学研究費補助金第3次対がん総合戦略研究事業「国民のがん情報不足感の解消に向けた「患者視点情報」のデータベース構築とその活用・影響に関する研究．2011年．
3）中山健夫．平成23年度厚生労働科学研究費補助金第3次対がん総合戦略研究事業「国民のがん情報不足感の解消に向けた「患者視点情報」のデータベース構築とその活用・影響に関する研究．2012年．
4）中山健夫．平成24年度厚生労働科学研究費補助金第3次対がん総合戦略研究事業「国民のがん情報不足感の解消に向けた「患者視点情報」のデータベース構築とその活用・影響に関する研究．2013年
5）Quinnlan Ross. Data Mining from an Al Perspective. Data Engineering, Proceedings.15th International Conference on. 1999.
6）Leoreiman. Charles. J.S.R.A. Olshen. Classification and Regression Trees. CHAPMAN&HALL/CRC. New York. 1998.
7）Breiman. L. and Friedman. J. Predicting Multivariate Responses in Multiple Linear Regression" (with discussion). J. Roy. Statist. Soc. B 59, 3. 1997.
8）Breiman. L. Random Forests, Machine Learning, 45, pp.5-23. 2001.
9）DIPEx-Japan ホームページ．http://www.dipex-j.org/（214．4 .14閲覧）
10）松本裕治他．形態素解析システム「茶筌」version2.2.7使用説明書．奈良先端科学技術大学大学院松本研究室発行．2001．
11）http://cl.aist-nara.ac.jp/lab/nlt/chasen.html（2014．4 .14閲覧）
12）Marshall. H. Bechker. et al. The health belief model and prediction of dietary compliance: a field experience. Journal of Health and Social Behavior 18. 348-366. 1977.
13）David S. Gochman ed. Health Behavior: Emerging Research Perspectives. Springer, 1988.
14）Sato RS. Beppu H. Iba N. Sawada. A. The meaning of life prognosis disclosure for Japanese cancer patients: a qualitative study of patients' narratives. Chronic Illness 2012.

Abstract

The purpose of this study was to clarify the possibility of predicting screening actions by quadrat analysis using artificial intelligence, and from the text data of cancer screening based on the interviews of patients who had an experience of having a breast cancer and a prostate cancer in Japan. The data was created as part of a clinical cancer research project of the scientific research subsidies from the Ministry of Health, Labour and Welfare, and managed by DIPEx-Japan.

The subject of research was the data including 38 cases of breast cancer patients and 46 cases of prostate cancer patients which was obtained in a maximum variation sampling survey, arbitrary sampling with consideration of the disease and regional characteristics. The random forest method, which is collective machine learning utilizing artificial intelligence for text data, was adopted to determine the prediction performance using proper nouns that were estimated from a model created based on the gini coefficient.

As a result, the gini coefficient improved the differentiating performance regarding the presence or absence of screening visits, in the order of "supplements" and "PSA" in prostate cancer patients, and "mammo" and "edema" in breast cancer patients. In terms of the prediction performance of the model, it showed 46.7% of distinguishability in prostate cancer patients and 59.5% in breast cancer patients. The models generated by collective learning and machine learning using artificial intelligence were reversed between the two regarding the ratio of frequency of medical terminology and general terms.

Since it was possible to obtain such findings by using artificial intelligence in the quadrat analysis method, performing a data mining shortly after the start of interviews was suggested to have a possibility of making it easier to create structured questions in order to predict the health behaviour with more relevant information.

Key words : Predictive model of screening behavior, random forest machine learning, text mining, breast cancer experience, prostate cancer experience