DEIM Forum 2023 XX-Y

# 特定健診データに含まれる自由記述に着目した予後予測

大塚　皇輝† 　池之上辰義†† 　福間　真悟††† 　若宮　翔子† 　荒牧　英治†

† 奈良先端科学技術大学院大学 先端科学技術研究科情報領域　〒 630-0192 奈良県生駒市高山町 8916 番地の 5
†† 滋賀大学 データサイエンス AI 推進センター　〒 522-8522 滋賀県彦根市馬場 1 丁目 1-1
††† 京都大学 医学研究科 人間健康科学科系専攻　〒 606-8507 京都府京都市左京区聖護院川原町 53
E-mail: †{otsuka.koki.om2,wakamiya,aramaki}@is.naist.jp, ††saikawa@n-univ.ac.jp,
†††fukuma.shingo.3m@kyoto-u.ac.jp

**あらまし**　日本は少子高齢化の影響によって医療財源の逼迫が問題となっており，予防医療や実診療での適切な医療資源の配分が求められている．このような背景から，近年では機械学習を用いて患者の予後を予測することで，医療分野における意思決定を補助するような試みも登場してきた．一方で，機械学習はその予測過程がブラックボックスであり説明可能性という観点から医療分野での導入の障壁となっている．そこで本研究では自然言語の持つ説明可能性の高さに着目し，特定健診データから言語モデルを用いて予後の予測を試みた．特に，減量の予後予測に対する言語モデルの予測精度と説明可能性について検証を行った．

**キーワード**　自然言語処理，医療情報学，予後予測，解釈可能性，特定健診

## 1　はじめに

日本は少子高齢化や人口減，労働人口現象に伴う財政難など様々な問題を抱えている．これらに起因して日本の税収に占める社会保険料の増加が問題となっており，医療の人的・物的資源の効率的な利用が課題となっている．特に，日本では悪性新生物（がん）や心疾患といった一般的には生活習慣に起因する疾患が死因の上位を占めている [1]．これに対して，日本では 2008 年より生活習慣病の予防を目的として 40 歳から 75 歳までを対象に特定保険審査，及び生活習慣病リスクの高い患者に対しては特定保健指導を実施している．しかし，指導を行なっても生活習慣が改善されない場合も多く，効果を疑問視する声もある．例えば運動や禁酒を薦められ，しばらく努力しても諦めてしまうといった事例も多い．どのような患者にどのように指導を行うかは，保健師・栄養士などの指導員の経験によるところが多く，指導の質の向上と均質化が求められる．

このような中，近年急速に発達している機械学習技術に注目し，この技術を医療分野に応用するような研究も進んできた．特に，患者予後を予測することができれば，予防医療や実診療での医療資源の最適な利用に繋げることが可能となる．しかし，機械学習によって導き出された予測結果は，解釈性に乏しいという問題がある．基本的に機械学習モデルの予測過程はブラックボックスとなっており，人間が解釈を行うことが難しい．医療分野では治療をする上で解釈性の乏しい機械学習モデルの結果を信頼することは難しく，このことが医療分野における機械学習を用いた予測モデルの導入の障壁となっている [2]．

そこで今回は自然言語を有効に利用することで，医療現場における説明可能性という障壁を取り除けないかと考えた．医療データには様々な種類のデータが含まれており，検査値のような数値データに加え，医師の診断所見といった自由記述が多く含まれている．しかし，従来ではこれらの自然言語を機械学習モデルで扱うことが難しく，医学研究の材料としてあまり用いられてこなかった．数値データ単体を人間が解釈するためには十分な経験が必要とされるが，自然言語は比較的人間には解釈しやすいため，この特性を有効活用することで医療分野における機械学習の導入の一助となる．

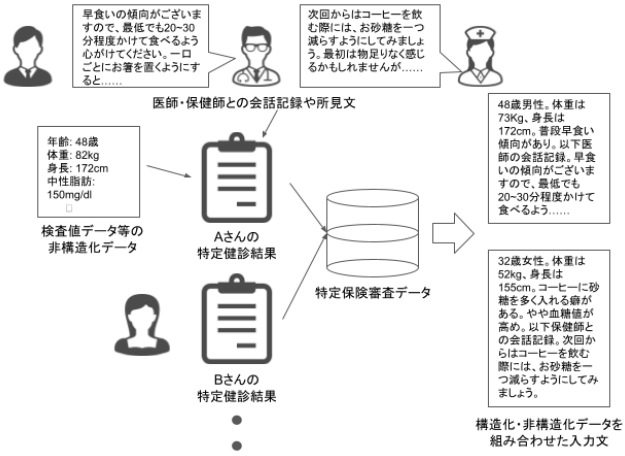


図 1　本研究の概要: 特定健診データに含まれる様々な種類のデータを言語モデルが解釈できるような自然言語に変換し，予後を予測する．

このような背景から，本研究では自然言語の持つ解釈性に注目した予後の予測, 及びその根拠となる情報を提示することにより，解釈可能性の検証を行った．予後予測の研究では検査値等の数値データや，カテゴリカルな検査データを入力として用いることが一般的となっている．しかし，このようなデータを入力に用いた場合，どの項目が予測結果に寄与したのかを判定

することしか出来ず，その数値に対する解釈はそれを見た医師の経験に左右される場合がある．一方で，自然言語の場合数値データよりもコンテキストが含まれているため，解釈を行う上では有効に作用する．加えて，電子カルテに代表される診断記録には医師の自由記述による所見も多く含まれているものの，この部分の利活用はあまり進んでいない．

本研究では，特定保険審査データと言語モデルを用いて，予後が良好か不良かを当てる分類問題を設定し，予測性能の検証をした．本研究で定義する予後とは，次回健診時までの腹囲と体重の差分から判定をする．具体的には，予後良好は腹囲が 2cm 以上減少し，同時に体重が 2kg 以上減少している場合とし，予後不良はそれ以外の状態としている．加えて，分類結果の判断根拠となる自然言語部分を可視化することで，診断の判断根拠として使用できるかどうか検証を行った．

## 2 関連研究

### 2.1 健診データを用いた研究

健康診断は 1 年に 1 回定期的に実施されることから時系列的な変化を追いやすい．そのため，このデータを軸にして様々な他媒体のデータと組み合わせた研究が盛んに行われている．

恒川ら [3] は生活習慣病の発症を予測するにあたり，健診データとレセプトを利用した．レセプトとは，医療機関が医療保険者に対して行った検査や診断，処方した薬剤等の情報が記された診療報酬明細書のことである．彼らの研究では，生活習慣病の発症を予測するにあたり，1 年に 1 回しかない健診データだけを用いた場合，生活習慣病になった時期の詳細が不明であるためレセプトを同時に使用した．作成されたデータには数値データや医師の自由記述など様々な形式のデータが含まれることになる．これらを解析した結果，生活習慣病の中でも糖尿病が識別しやすい病気であることが発見され，HbA1c や糖代謝判定といった項目が予測に大きく寄与していることがわかった．一方で，この研究では心電図の結果等に関する医師の自由記述項目がデータに含まれていたが，自然言語の処理を組み込むことができず今後の課題として紹介されている．

大場ら [4] も同様に健康診断データを利用して，生活習慣病の中でも特に糖尿病を予測するという研究を行った．彼らの研究では，予測精度だけではなく，医療分野での使用を想定し解釈可能性といった点にも注目している．健診データに加えて質問票への回答から予測に寄与するデータを明らかにしている．これを調査するにあたり，構築したモデルの入力として使われている属性をランダムに置き換えて変化を計測する Permutation importance [5] と，摂動を与えてその前後の予測結果の違いを分析する Sensitivity analysis といった手法を利用している．前者は変化が大きかった場合，その属性が予測結果に与える影響が大きかったと判定することができるが，その解釈を行うことはできない．一方で，後者は摂動を与えた際の前後の変化を観察することで，どの程度の変化がどのような予測結果に紐づいているか解釈することが可能となる．

### 2.2 マルチモーダルなデータ活用

健診データのように一度に複数項目の検査を実施すると，検査値データや心電図等の時系列データ，医師の所見などの自然言語等様々な種類のデータが作成される．それぞれ検査目的が異なるため，どの項目を用いて予測を行うべきかは医師などのドメインエキスパートによって選ばれる．そこで，例えば，モデルが数値データだけではなく，医師が診断時に感じ取った患者の状態等を記した自由記述データを同時に処理することができれば，より高精度な予測が可能になる．こういったことから，Agarwal ら [6] はマルチモーダルなデータを一つのモデルで処理することで，COVID-19 の患者の予後を予測する研究を行った．具体的には，一定期間後にまだ入院しているかどうか，人工呼吸器を装着することになるかどうかの 2 点を予測項目とした．これらの項目を予測する際に，性別や身長といったデータから飲酒喫煙の状態，検査値等の連続データを一つのベクトルとして埋め込むことでモデルの入力として利用している．

## 3 データセット

### 3.1 特定保険審査・特定保健指導データ

日本では生活習慣の変化により，高血圧症や糖尿病等のいわゆる生活習慣病の有病者が増加している．日本人の死因上位における悪性新生物や心疾患の原因には生活習慣病があるとされ，これらの対策が急務となっている．このようなことから日本は平成 20 年 4 月より生活習慣病の予防を目的として，年に一度 40 歳から 75 歳までを対象に内臓脂肪肥満に着目した特定保険審査を実施している．これには身体計測や血液検査に加え，質問票を用いた問診も含まれる．この検査で生活習慣病の発症リスクが高く，生活習慣の改善で予防ができると判断された場合，医師や保健師による特定保健指導により，生活習慣の改善に取り組むことになる．今回は，全国土木建築国民健康保険組合加入者に対して，2013 年 3 月から 2017 年 3 月の間に行われた特定保健師指導の電子的報告記録から抽出した 28954 件を，データセットとして用いた．

このデータには「年齢」や「身長」といった 54 種類の構造化データと，「既往歴」や「生活習慣病の改善状況」といった 11 種類の非構造化データが含まれている．表 1 に含まれている項目の一部を示す．

### 3.2 予測ラベルの作成

3.1 節で示した通り，特定保険審査は生活習慣病の予防を目的とし，内臓脂肪型肥満に着目した健康診断となっている．そのため，今回は内臓脂肪の数値に関係すると考えられる腹囲と体重に注目した．なお，これら 2 つの項目については，次回健診時の数値も入力されていた．そのため，これらの数値を利用し，次回健診から腹囲が 2cm 以上減少しかつ，体重が 2kg 以上減少していた場合予後が良好であると判定することとした．

### 3.3 不均衡データの扱い

表 2 に示す通り予測対象となる予後が良好となっているデー

表 1　特定健診データに含まれる項目の一部．下線は入力文章の作成に用いた項目を表す．

| 種別 | 項目名 |
|---|---|
| 構造化データ | 患者 ID，喫煙有無，年齢，性別，生年月日，<br>歩行速度，貧血，飲酒頻度，一年間の体重変化<br>30 分以上の運動習慣，腹囲，BMI，身長，体重<br>収縮期・拡張期血圧，GOT(AST)，HbA1c，<br>LDL コレステロール，HDL コレステロール<br>中性脂肪，空腹時血糖，尿糖，尿蛋白<br>血清クレアチニン，次回健診時体重，<br>次回健診時腹囲，診断回数，ALT(GPT) |
| 非構造化データ | メタボリックシンドローム判定，保健指導レベル，<br>既往歴，自覚症状，自覚症状初見，<br>食べ方，飲酒量，保健指導時の会話記録，<br>飲酒習慣詳細，睡眠習慣詳細，生活習慣改善意思 |

タが 2698 件，不良となっているデータが 28954 件となっており大きな偏りが生じている．このまま言語モデルを学習すると予後不良となっている多数派の影響を強く受けることでモデルの出力がほぼ予後不良となるように学習してしまう危険性がある．そのため今回はデータのダウンサンプリングを実施した．予後不良となっているデータからランダムに予後良好となっている件数と同じ 2698 件を抽出して用いることにした．これにより予後良好が 2698 件，予後不良が 2698 件で同数となるデータセットを作成した．

## 4　手　　法

### 4.1　モ　デ　ル

予後の良好・不良を予測する分類問題を解くに当たって，今回は Bidirectional Encoder Representations from Transformers (BERT) [7] をベースとした 2 つの言語モデルを用いて実験を行った．1 つ目は 2019 年 9 月までの日本語版 Wikipedia に含まれる，約 1700 万行の文書を用いて事前学習した bert-base-japanese-whole-word-masking（以下 Tohoku-BERT）[8]，2 つ目は東京大学附属病院の電子カルテに保存されていた約 1 億 2000 万行の文書をコーパスとして事前学習した UTH-BERT-BASE-512-WWM (以下 UTH-BERT) [9] を用いた．これらのモデルは両方とも日本語を学習したモデルであるが，前者が汎用的な知識を獲得しているのに対し，後者は医療分野に特化した知識を獲得している．今回使用するデータに含まれる自由記述部分には医師・保健師と患者による対話が収められた部分が多く存在している．そのため医療知識をあまり持たない患者向けに，噛み砕いた説明になっていることが予想されたため，医療特化のモデルだけではなく，汎用的な知識を保持した日本語モデルも用いることとした．

構造化データを日本語に変換して結合した入力文

これは 4 回目の診察です．性別は男性．GPT(ALT) は正常．HbA1c の値は要注意．

非構造化データを日本語に変換して結合した入力文

先日はご多忙の中，お時間を頂きましてありがとうございました．〇〇〇〇〇〇です．足の痛みが和らいだこと，痛風の治療と服薬は医師の指示どおり行えていることを伺い，安心しました．今回のお電話では，...

構造化・非構造化データを結合した入力文

これは 4 回目の診察です．性別は男性．GPT(ALT) は正常．HbA1c の値は要注意．先日はご多忙の中，お時間を頂きましてありがとうございました．〇〇〇〇〇〇です．足の痛みが和らいだこと，痛風の治療と服薬は医師の指示どおり行えていることを伺い，安心しました．今回のお電話では，...

### 4.2　入力文章の作成

今回は言語モデルを用いて予測を行うことから，入力は自然言語による記述となるため，データセット内の特に自由記述部分の項目を主に扱うこととした．主な自由記述項目としては「生活改善指導時の発話」「継続配慮に関する申し送り」「飲料習慣」「睡眠習慣」「雑記」の 5 項目となっていたため，まずこの文章同士を繋げて一つの入力文章とした．加えて構造化データ部分に関しても予測結果に寄与すると考えられる項目が複数含まれていることから，これらの項目に関しても入力として用いることとした．図 2 はこのようにして作成された入力文章の一例である．特定保険審査データには 54 種類の構造化データが含まれていることは 3.1 節で紹介した通りである．これらの項目すべてを入力文章として変換して用いることは，今回使用するモデルの入力トークン数の最大が 512 トークンとなっていることから困難であると考えた．そのためこの 54 項目のうち予測に寄与すると考えられる「性別」「ALT(GPT)」「HbA1c」「中性脂肪」「診断回数」の 5 項目を入力文章として追加することとした．この中で ALT(GPT) は肝臓脂肪に関連する数値である．HbA1c はヘモグロビンのうち糖化したものの割合を示しており，主に糖尿病に関わる数値となっている．これらの数値を単純に文章化すると「性別は男性．ALT(GPT) は 50IU/L．HbA1c は 4.3 %．中性脂肪は 69mg/dl．診断回数は 2 回」といったなる一方で，「ALT(GPT)」「HbA1c」「中性脂肪」に関してはその数値が異常なのか判断がつかない．そのため，今回はこの 3 項目に関しては基準値を調べ，ALT(GPT) の数値が 4 未満であれば「低い」，HbA1c の数値が 5.8%であれば「正常」，中性脂肪の数値が 149mg/dl より大きければ「多い」という形で変換して文章を作成した．作成した文章とその数を表 2 に示す．また，作成した文章のトークン数を調べた結果を図 3 に示す．これによりデータのほとんどがモデルの入力トークンの最大値である 512 トークン内に収まっていることを確認した．

表 2　カテゴリを一定の基準でグルーピングした際の基準と該当数

| 項目 | | 該当数 |
|---|---|---|
| 予後 | | |
| | 良好 | 26356 |
| | 不良 | 2598 |
| 性別 | | |
| | 男性 | 27841 |
| | 女性 | 1113 |
| GPT(ALT) | | |
| | 低い ($< 4IU/L$) | 1 |
| | 正常 | 23546 |
| | 高い ($44IU/L <$) | 5407 |
| HbA1c | | |
| | 正常 ($< 5.6$ %) | 12532 |
| | 要注意 | 13418 |
| | 極めて高い (6.4 % <) | 3002 |
| | 測定不能 | 2 |
| 中性脂肪 | | |
| | 少ない ($< 30mg/dl$) | 6 |
| | 正常 | 15175 |
| | 多い ($149mg/dl <$) | 13773 |

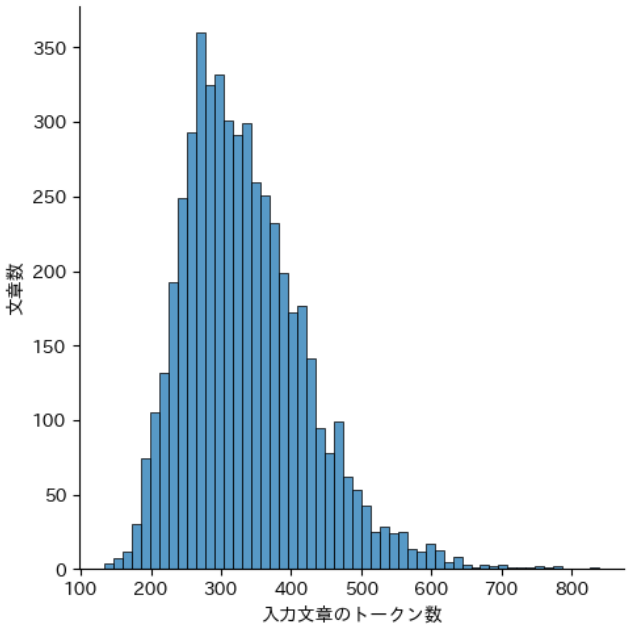


図 2　作成した入力文数とトークン数の分布

## 5　実　　験

### 5.1　設　　定

本研究では前回健診時からの患者予後が良好か不良かを判定する分類問題を設定した．データに関しては予後の数が不均衡になっていたため，これを解消するためにダウンサンプリングを実施し，最終的に 5396 件のデータを用いて実験を行った．モデルに関しては一般テキストを用いて事前学習した Tohoku-BERT と，医療分野のテキストで事前学習した UTH-BERT を用いた．入力文章に関しては構造化データのみを文章化したもの，自由記述欄のみを結合して文章化したもの，これら 2 つの文章を結合させて一つの文章にしたものの 3 種類の入力文章を作成した．評価指標に関しては分類タスクということを考慮し F 値，AUROC，Accuracy の 3 種類を用いた．加えて予後判定時に用いる閾値を 0.1 ごとに区切って感度と特異度を計算した．感度とは実際に予測対象の疾病に罹患している人のうち予測結果が罹患しているとされた人の割合を示し，特異度とは予測対象の疾病に罹患していない人のうち予測結果が罹患していないとされた人の割合を示している．本研究の場合は，予後が良好となっている患者に対して過剰な医療リソースの投下を抑制するといった状態を想定している．そのため予後良好のラベルがつけられた人のうち予測結果が予後良好となっている人の割合を感度，予後不良のラベルがつけられた人のうち予測結果が予後不良となっている人の割合を特異度として計算している．実験のハイパーパラメータとしてはバッチサイズは 32, エポック数は 1000, アーリーストップは 50, 学習率は 2e-5 とし，消費した GPU のメモリは約 29GB, 約 100epoch で停止し, 計算時間は約 2 3 時間ほど要した．

### 5.2　結　　果

表 2 に実験結果を示す．まず入力データセット間における精度は，構造化データのみ，自由記述欄のみ，構造化データ+自由記述の順に全体として精度が向上することが確認できた．構造化データのみを文章化したものを入力として用いた場合は UTH-BERT のほうが性能が高く，自由記述欄のみを結合して文章化したものを入力として用いた場合は Tohoku-BERT のほうが精度が高くなった．検査値データが含まれた文章と自由記述欄を組み合わせて作成した文章の場合，F 値に関しては UTH-BERT のほうが精度が高く，AUROC と Accuracy に関しては Tohoku-BERT のほうが精度が高いという結果となった．

表 4 は感度・特異度の結果を表している．医療資源の最適化といった場面で過剰な医療リソースを投下しすぎないということを目的とした場合, 閾値を 0.4 とした時の感度が Tohoku-BERT と UTH-BERT 共にが 85%ほどと十分な精度となっている．

図 4 は入力文章に Attention を表示して可視化した結果である．Attention が全体的にかかっており，ひと目見て注目するべき点がどこかを判定することは難しい．

### 5.3　考　　察

#### 5.3.1　モデルの特性について

モデル別の性能を比べてみると，検査値データのみを入力文章とした場合は UTH-BERT のほうが全体的な精度が高くなっている．一方で，自由記述部分のみを入力文章とした場合は Tohoku-BERT のほうが全体的な精度が高くなっている．UTH-BERT は医療テキストを用いて事前学習されていること

表 3　入力とモデル別に測定した性能の一覧．構造化データと非構造化データを組み合わせて作成した文章を入力に用いた際に一番精度が高くなる．

| | 構造化データのみ | | 非構造化データのみ | | 構造化+非構造化データ | |
|---|---|---|---|---|---|---|
| 評価指標 | Tohoku BERT | UTH-BERT | Tohoku BERT | UTH-BERT | Tohoku BERT | UTH-BERT |
| F 値 | 0.630 | 0.629 | 0.646 | 0.015 | 0.658 | **0.683** |
| AUROC | 0.428 | 0.621 | 0.644 | 0.523 | **0.662** | 0.642 |
| Accuracy | 0.478 | 0.605 | 0.619 | 0.501 | **0.630** | 0.615 |

表 4　閾値をずらして感度と特異度を計算した結果
表 5　予後ラベル判別の閾値をずらして，感度と特異度を計算した結果

| | 感度 | | 特異度 | |
|---|---|---|---|---|
| 閾値 | Tohoku BERT | UTH-BERT | Tohoku BERT | UTH-BERT |
| 0.0 | 1 | 1 | 0 | 0 |
| 0.1 | 0.9769 | 0.9961 | 0.0961 | 0.0653 |
| 0.2 | 0.9269 | 0.9961 | 0.2461 | 0.1038 |
| 0.3 | 0.9038 | 0.9615 | 0.3 | 0.1807 |
| 0.4 | 0.8576 | 0.9 | 0.3846 | 0.3346 |
| 0.5 | 0.7115 | 0.8307 | 0.55 | 0.4 |
| 0.6 | 0.2461 | 0.5769 | 0.8384 | 0.5961 |
| 0.7 | 0.0038 | 0.1807 | 1 | 0.8884 |
| 0.8 | 0 | 0 | 1 | 1 |
| 0.9 | 0 | 0 | 1 | 1 |
| 1.0 | 0 | 0 | 1 | 1 |

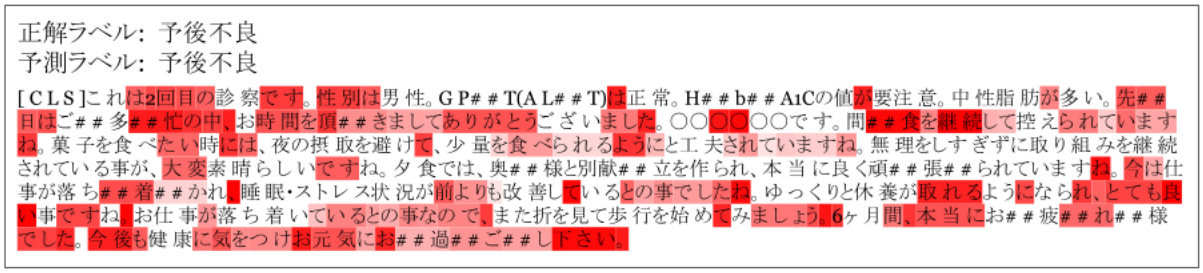


図 3　入力文章に対する予測根拠部分の Attention 可視化．

から，検査値データに関する認識精度が高かったと考えられる．一方で，自由記述欄の内容は医師・保健師と患者の会話時の文章が多く含まれており，非医療従事者向けの語彙が多く含まれている．そのため，一般的な語彙を多く含む Tohoku-BERT の方が認識できる単語が多く含まれていることから，UTH-BERT よりも Tohoku-BERT の方が精度が高くなったと考えられる．検査値と自由記述文の両方を含む内容を入力とした場合，入力文に含まれる情報としては自由記述欄の情報の方が多く含まれるようになっている．そのため Tohoku-BERT の方が若干精度が高くなったと考えられる．

**5.3.2**　分類性能について

実験では入力文を検査値と自由記述文から作成したとき，AUROC が 0.662，Accuracy が 0.630 程度にとどまっている．医療現場で使用することを想定すると，精度改善の余地が十分に残されている．

まず，本研究では，使用するデータセットを大きくダウンサンプリングしているため，予後不良ラベルが付与された未使用データが多く存在する．これを有効に利用するため，予後ラベルの出現割合が等しくなるようにリバランスしたデータセットを複数作成し，データセットごとにモデルを訓練した後，そのモデルの出力結果を多数決する方法が考えられる [10]．これにより，全データをモデルに投入することが可能となり，より広範囲な知識が獲得できると考えられる．一方で，今回の言語モデルのようにメモリを多く使用する場合，データセットごとのモデルを同時に学習するというのは難しい．そのため，一つのデータセットとモデルごとの結果を保存し，すべての計算が終わった後に集計処理を行うといった実装上の工夫が必要となる．

また不均衡データを，基準とする項目の個数が均衡するようにデータセットを改変して学習した後，不均衡データ全体を使って学習を行うと精度が向上するといった研究結果が存在する [11]．今回の実験では，データを均衡状態にしたため，作成したモデルを全体のデータを使ってもう一度学習することで，

精度が向上する可能性がある．

加えて，今回は，体重 2kg 以上減少かつ腹囲 2cm 以上減少した場合に予後良好とみなしたが，この条件が厳しかったという問題が考えられる．全データのうち，この条件を達成していたのは約 9%であったため，今回はこの数値に対して大きくダウンサンプリングを実施することになり，十分な精度を出すことができなかった．これに関しては，実際の現場で予後が良好と判断される人の割合と相違が無いのか調査する必要があると同時に，実験を実施する上では，条件を緩和する方法も考えられる．例えば今回の予後良好ラベルの設定を体重 1kg 以上減少，腹囲 1cm 以上減少とすることにより予後良好判定となるデータを増やすことができる．その状態での精度まず確認した上で，より厳しい条件の精度を確認するといった方法も考えられる．

**5.3.3** 予測確率の分布について

本研究ではデータをダウンサンプリングして予後の良好と不良が同数になったものをデータセットとして使用している．それに伴って閾値を 0.5 としてた場合にモデルの出力も予後の良好と不良の出現割合がおよそ 1 対 1 になるようになっている．しかし現実には予後が良好となる人の割合は少なく，今回のデータセットの場合約 9%しか存在していない．仮に本研究で作成したモデルが実際の医療現場で使えるだけの精度が出た場合でも，実験に使ったデータセットの分布が現実の患者予後の分布と異なるため，今回の場合では予後良好となる確率が高くなってしまう．この場合予後が不良の患者を予後良好と誤診する場合が想定され，重大な医療過誤につながる恐れがある．そのため予測確率の補正が必要になってくる．この場合現場で使用するといった際にはデータセット上の分布と現実の分布の乖離に注意し，Isotonic Regression [12] などの手法で現実のデータに適応するように確率補正を行うという処理が必要となってくる．

**5.3.4** 説明可能性について

本研究では従来の手法では確認することが難しい予測結果の根拠という部分に着目した．そこで可視化結果に注目すると，全体的に Attention がかかっているようになっており，特に際立った根拠部分を見つけるというのは難しい．今回は Attention 部分の可視化を行ったが，他には Integrated Gradients [13] といった手法が存在している．一方で Attention がかかっている部分が根拠として解釈してもよいのかという問題が存在しており，自然言語を用いた解釈性といった部分にはまだ検討の余地が残されている．

**5.3.5** 限　　界

特定健診の予後予測は経験を積んだ指導員にとっても難しい課題となっている．生活習慣の改善が求められるため，血糖値等の数値データよりも本人の家庭環境や改善への意思といったデータのほうが，予測する上では大きく寄与すると考えられる．例えば日々の仕事に追われておりほとんど家に帰宅しないといった人に対して，食生活の改善を促したとしても，それを実行するのが困難である場合が多いと考えられる．しかし従来の数値データを用いた手法ではこのコンテキストを把握することは難しく，そこに医師や保健師が診察を通じて記した自由記述データを有効に活用する価値は多く残されている．一方で本研究の結果の通り精度や解釈性には多くの課題が残されており，今後どのように活用していくか検討の余地が残されている．

**5.3.6** 今後の展望

近年 ChatGPT に代表されるように，BERT よりも多くのパラメータを持つ言語モデルが登場してきた．医療分野に特化したものでは Med-PaLM [14] というモデルが登場している．モデル自体はまだ一般公開されているものでは無いが，既に米国の医師国家試験と同等レベルとされる MedQA タスクにおいて約 86.5%もの精度を達成したと報告されている．このような医療分野に特化したより大規模なモデルは日々開発させており，自然言語を用いた病院内データの利活用はより進んでいくと考えられる．

このような巨大な言語モデルが発達していくに連れて，人間がこれらを扱う方法についても変わってくる．本研究で用いた BERT のようなモデルは，一定のタスクを設定してそれに対してモデルを学習させて使用するといった形が一般的であった．そのため入力は文章，出力は数値といった形になるとその数値の算出過程を明らかにすることが難しいため解釈方法といった別の問題も出てきていた．一方で近年の巨大な言語モデルは，一つのモデルをあらゆる大量のデータで学習することで，そのモデルですべてのタスクを解くという方向に変わってきている．このようなモデルは入力，出力ともに文章となっているため，入力時の命令に「出力時には予測根拠も出力すること」といったようなプロンプトを入力しておくことでモデルの予測根拠のようなものを同時に出力することができる．また，ChatGPT のようにチャット形式で使用することで人間が補完しきれていない知識を取り出すことが可能となっている．そのため人間側が機械の出力結果を使用できる程度のタスクに落とし込むことさえできれば，解釈性といった問題を無視することができる．例えば出力に対する信頼性が低くとも，人間がその出力に対して裏取り調査を行った上で利用する等が考えられる．このような変化が起こる中で，言語モデルの進化に伴い人間側もその使用方法を改善することで導入の障壁を取り除くことができるのではと考えられる．

## 6 おわりに

本研究では自然言語の持つ解釈性に着目し，特定健診データを使用して患者の予後予測を行った．データには構造化データも含まれていたため，これらを活用するべく 3 種類の文章を作成し，言語モデルの入力文章とた．予測ラベルに関しては次回健診時の体重と腹囲の変化から作成した．言語モデルは一般分野のテキストで事前学習済みの Tohoku-BERT と，医療分野のテキストで事前学習済みの UTH-BERT の 2 種類を使用した．これらを組み合わせた 6 種類の条件から予後の予測精度を調べた結果，約 63%程度の正解率となることが判明した．加えて医療分野で使用するためには説明可能性が重要になってくるため，今回はこの予測モデルの Attention を可視化することで根拠部分の判定が可能か調査を行った．その結果，現段階

では Attention のかかった部分を見ることで判断の根拠となる部分を判別することは難しかったが，一部診断回数には比較的 Attention が強くかかっていることが判明し改良次第では実応用可能なのではないかと考えられる．

## 7 謝　　辞

本研究の内容の一部は戦略的創造研究推進事業（CREST），及び JSPS 科研費 JP 19K10620 の助成を受けたものである．また本研究に使用したデータは，SOMPO ヘルスサポート株式会社および全国土木建築国民健康保険組合より提供を受けたものである．なお，本研究は京都大学医の倫理委員会（R0817-3）で承認を受けている．

## 文　　献

[1] 令和 4 年 (2022) 人口動態統計月報年計 (概数）の概況. p. 10, 2022.
[2] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems, Vol. 32, No. 11, pp. 4793–4813, 2021.
[3] 荒木雅弘 新谷元司 吉川昌孝恒川充. 健診データを用いた生活習慣病の発症予測. 2019.
[4] 讃岐勝 我妻ゆき子大場勇貴. 健康診断データを用いた疾患予測における解釈可能なモデルの構築. 2020.
[5] Oliver Sander Thomas Lengauer André Altmann, Laura Tološi. Permutation importance: a corrected feature importance measure. Bioinformatics, Vol. 26, No. 10, pp. 1340–1347, 2010.
[6] Sutanay Choudhury Sindhu Tipirneni Pritam Mukherjee Colby Ham Suzanne Tamang Matthew Baker et al Agarwal, Khushbu. Preparing for the next pandemic via transfer learning from existing diseases with hierarchical multimodal bert: a study on covid-19 outcome prediction. Scientific Reports, Vol. 12, , 2022.
[7] Ming-Wei Chang Kenton Lee Devlin, Jacob and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.
[8] M Suzuki. cl-tohoku/bert-japanese: Bert models for japanese text. Accessed: 3-Nov-2023.
[9] Daisaku Shibata Emiko Shinohara Eiji Aramaki Kawazoe, Yoshimasa and Kazuhiko Ohe. A clinical specific bert developed using a huge japanese clinical text corpus. PLoS One, Vol. 16, No. 11, 2021.
[10] Kevin Small Carla E. Brodley Wallace, Byron C. and Thomas A. Trikalinos. Class imbalance, redux. In 2011 IEEE 11th international conference on data mining, pp. 754–763, 2011.
[11] Simon Wouters Ruben Cartuyvels Erfan Ghadery Malik, Farjad and Marie-Francine Moens. Two-phase training mitigates class imbalance for camera trap image classification with cnns. p. 10, 2021.
[12] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. pp. 625–632, 2005.
[13] Ankur Taly Sundararajan, Mukund and Qiqi Yan. Axiomatic attribution for deep networks. PMLR, pp. 3319–3328, 2017.
[14] Tao Tu Juraj Gottweis Rory Sayres Ellery Wulczyn Le Hou Kevin Clark et al Singhal, Karan. Towards expert-level medical question answering with large language models. 2023.