異種医療データの融合による医療需要予測手法の検討

吉本　廣雅† 　満武　巨裕†† 　合田　和生†

† 東京大学 生産技術研究所　〒 153-–8505 東京都目黒区駒場 4-–6-–1
†† 医療経済研究機構　〒 105–0001 東京都港区虎ノ門 1–21–19 東急虎ノ門ビル 3F
E-mail: †{yoshimoto,kgoda}@tkl.iis.u-tokyo.ac.jp, ††mitsutake@ihep.jp

あらまし　蓄積したデータから将来が予測できれば，それは意思決定の支援になる．本稿は医療データから将来の医療需要を個人単位で予測する問題を考え，異種の医療データを融合することで予測精度を高める手法について検討を行う．融合するデータは，医療レセプトデータと特定健診データである．前者は，病院・薬局で医療保険を利用した際の明細書（レセプト）であり，診察・治療・薬の処方などの医療行為とその医療費からなる縦断的データである．後者は，特定健診を受診した際の体重や血圧などの横断的データである．実験では 7 年間，約 241 万人のデータを用いて 1562 項目の医療需要を個人毎に予測し，結果，提案手法により 1035 項目 (66%) で予測精度が向上することを実証した．本稿は，精度向上のメカニズムを分析するとともに，医療需要予測のための最適なデータ融合方法について考察をまとめる．

キーワード　医療需要予測，機械学習，医療レセプト，特定健診

1　は じ め に

蓄積したデータの活用法の一つは将来の予測である．日本では厚生労働省が健康医療分野のデータをナショナルデータベース（NDB) として蓄積している．NDB には，国民の一人一人が病院や薬局で保健医療を利用した際の診療報酬明細書（以下，レセプト）や健康診断の結果（以下，特定健診）などが全て蓄積されている．これはまさに健康医療のビッグデータであり，研究者や医療従事者が将来をより正確に予測し，医療を改善するための貴重な情報源となりうる [1,2]．我が国が直面している問題，少子高齢化，人口減，財政危機を考えると，今後社会が劇的に変化することは明らかである．その変化の一つは医療需要である．住民の一人一人が将来どの種類の医療や介護サービスをどれだけ必要とするか，その質と量を正確に予測できれば，人的・医療的資源を最大限活用することで，医療サービスの効率化や質の向上，患者のアウトカム改善など，医療・介護・行政への貢献が期待できる [3–6] ．そのためには，NDB のようなビッグデータを活用して将来の医療需要を予測する情報処理技術の開発が必要不可欠である．

ここで医療需要とデータの関係は次のようになる．まず，レセプトは医療サービスの提供側である医療機関が作成するデータであり，住民に提供した医療サービスの記録である．具体的には， レセプトには投薬した薬の種類や回数などがその費用（点数）とともに正確に記録されている [1]．一方で，特定健診は住民が健康診断を任意で受診した際に記録されるデータであり，データは住民の健康状態に関する情報（血圧や心電図，喫煙の有無などの問診結果）となる [2]．ここで，レセプトのデータも特定健診のデータもどちらも直接は医療需要を捉えたデータとならない点が重要である．レセプトの本来の利用目的は医療機関の医療費の請求であり，特定健診の本来の目的は生活習慣病（メタボリックシンドローム，高血圧，糖尿病，脂質異常症など）の早期発見と早期対策である．つまりレセプトや特定健診のデータは医療需要と一対一に対応するとは限らず，医療需要を間接的に観測したデータとなっている．

本研究の目的は，これら医療需要を直接は観測できていない既存のビッグデータを活用することで，その背後にある医療需要を予測する手法を明らかにすることにある．そのアイデアは，医療需要を潜在変数と捉え，それがレセプトや特定健診のデータとして観測・記録される過程を考慮する点にある．以下，本稿は，レセプトとしての観測と特定健診としての観測の二つの観測を考え，その観測を統合することでより将来の医療需要を正確に個人単位で予測する手法について議論を行う．

本稿の構成は以下の通りである．2 節では関連研究について述べ，3 節で問題の定式化を行い，提案法について説明する．4 節と 5 は実験結果と考察であり，最後に 6 節で結論をまとめる．

2　データに基づいた医療需要予測

医療需要の予測では様々な手法が利用されている．AR モデルのような時系列分析を用いる手法 [7,8]，決定木 [9,10] や support vector machine（SVM）[11] などの機械学習を用いる手法 [6,12,13]，LSTM のようなディープラーニングを用いる手法 [14–16]，MCMC のようなベイズ推定を用いる手法 [17,18]，である．これら手法は，処理の枠組みが教師あり学習である点で共通している．すなわち，ある過去の時刻 t でのデータを教師データとし，予測器が t よりも過去の情報から時刻 t の値を正しく予測できるように予測器を訓練する処理となっている．

教師あり学習という観点で見ると，先行研究での議論は，予測器のアルゴリズムに関する議論と，予測器に与えるデータ処

理の方法の議論とに大別できる．

前者の議論としては，例えば，文献 [13] は機械学習の手法として 3 種類のアルゴリズム， gradient-boosting decision tree(GBDT) [9]，multilayer perceptron (MLP) [19]，SVM [11] を取り上げ，ベンチマークによる性能比較を行なっている．このように教師あり学習では，複数の手法を比較し，最も性能が良いアルゴリズムを選択する手法が広く用いられている．

また一般に教師あり学習は学習データを増やせば増やすほど予測精度の向上が期待できる．例えば文献 [16] はデータ数の不足による予測精度の低下を指摘している．その背景としては，医療に関する事象，傷病やそれに対する医療行為には膨大な組み合わせが存在する一方で，各事象の発生頻度は低いことが主因である．例えば，日本では白血病や癌のような診断がつく患者は 1 年間で 10 万人当たり 10 名程度，骨折で入院する患者は 10 万人あたりで 70 名程度である．このように個人単位で見ると医療需要の発生は低頻度である．結果，医療データはスパースなものになり，個人単位での医療需要の予測はスパース性に起因した難しさを抱えることになる．

データのスパース性の問題を解決する手法として，文献 [20] は学習側の処理を改善し，データのサンプリング粒度を複数用意しその中から最適な粒度を選ぶ手法を提案をしている．これは，スパースなデータから識別器にとって最適な解像度で特徴ベクトルを自動選択する，適応的な特徴量選択手法と解釈できる．

データの母数は増やさずにデータの中で利用可能なデータを増やす方法もある．例えば医療データは医療機関，医療サービス毎に異なるシステムで管理されており，データの名寄せの失敗により利用可能なデータが減少する問題がある．これに対して [21,22] は，時間方向で医療データを観測し個人を追跡する，名寄せの高度化手法を提案している．これにより利用可能データの割合が増え，データセットのサイズを大きくすることができる．

適応的な特徴量選択や，名寄せの高度化は，手元にある医療データに対して利用可能なデータの割合を増やす技術である．一方で，医療データが生成・記録される過程を考えると，そもそも手元にあるデータに全ての医療行為，医療需要が記録されているわけでは無い点も考慮すべきである．極端な例をあげれば，病気になっても病院に行かず医療サービスを受けなければデータは発生・記録されない．手元にある医療データが全ての事象であると捉えるのではなく，医療データは観測結果の一部であり，観測結果から事象の真の分布を推定するという考え方も必要である．

そこで本研究は事象の観測という観点から，教師あり学習のためのデータセットを生成する処理に焦点をあて，次に述べる新しい手法を提案する．

3　異種医療データの融合による医療需要予測

異種医療データとしてレセプトデータと特定健診データの二つを具体例として議論を行う．両者は個人情報保護の観点から仮名加工され，個人識別子 (ID1, ID2) が付与されている．本研究は，これら識別子を手がかりに，佐藤らの提案する vPID [21,22] を用いて個人を追跡，ある程度の範囲でデータの名寄せが可能であるものとする．

表 1: レセプトから抽出した 1562 項目の一部

種別	項目名
傷病名コード	骨折，心筋梗塞，慢性動脈閉塞症，耳鳴症，急性腸炎，アトピー性皮膚炎，アレルギー性鼻炎，胃癌，パーキンソン病，COVID-19
医薬品コード	ブドウ糖，ビタミン B1，ザイロリック錠，クラビット点眼薬，ダイアート錠，リクシアナ錠，アスベリン錠，アセトアミノフェン，アクアチム軟膏，フリバス錠，ビオフェルミン錠，ゲンタシン，ロキソニンパップ 100mg，リピトール錠，葛根湯，タチオン錠
診療行為コード	点滴注射，静脈内注射，インフルエンザウイルス抗原定性，耳処置，LDL-コレステロール，超音波検査，CT 撮影，初診料，往診料，夜間・早朝加算，皮膚科特定疾患指定管理料
特定器材コード	画像記録用フィルム，酸素ボンベ，皮膚欠損用創傷被覆材，オルソパントモ型，血管造影用ガイドワイヤー，胃管カテーテル，レジン歯臼歯用
歯科診療コード	刺繍基本治療，う蝕歯即時充填形成，咬合採得，有床義歯修理，抜髄，処方料，歯科初診料

3.1　レセプトデータ

レセプトデータは，ある住民 vPID について医療費の明細を月単位で集計したデータであり，医療費，医療費の根拠となる医療サービス（診療や調剤），処方した薬剤，使用した医療機材の単価とその数量，などが記録されている．医療サービスや薬剤には専用のコード体系がありユニークな識別子が付与されている．その識別子は数万種類あり，レセプトには識別子とそのサービスが発生した日付が共に記録されている．

本研究は，これらレセプトに含まれる情報を以下の手順で Bag-of-Words 表現（BoW) に変換する．まず各識別子の出現頻度を調べ，出現頻度が極端に低いもの (年間 10 万人あたり 1 件未満) は除外する．次に，医薬品はジェネリック医薬品のように同一成分を有する場合でも販売元により異なる商品名，医薬品コードが付与されている．そこで有効成分が同じ医薬品には同一識別子を付与する名寄せ処理を行う．最後に，過去 3 年間のタイムウインドウを設け，タイムウインドウ内でのサービスや薬剤の利用の有無を集計する．

以上の処理で，頻出する薬剤や医療サービスを 1562 項目に絞り絞り込む．表 1 に 1562 項目の一部を示す．項目は，医師の診断に基づいた傷病名，処方した医薬品，治療・診療で行った行為，その際に使用した医療機材・医療機器，歯科の診療行為，で構成されている．本稿は，これら 1562 項目の利用の有無 (0 or 1)，年齢，性別の情報からなる 1564 次元のベクトルとして $\boldsymbol{r}$ を定義し，その集合をレセプトデータ群 R と定義する．

表 2: 特定健診の 85 項目の一部

種別	項目名・質問項目
数量変数	身長，体重，BMI，腹囲，収縮期血圧，拡張期血圧，中性脂肪，HDL コレステロール，LDL コレステロール，non-HDL コレステレロール，GOT(AST)，GPT(ALT)，γ-GT(γ-GTP)，随時血糖，空腹時血糖，HbA1c，赤血球数，心電図，眼底検査
カテゴリ変数	既往歴（脳血管，心血管，腎不全・人工透析），服薬の有無（血圧，血糖，脂質），医師から貧血と言われたことがある，ほぼ同じ年齢の同棲と比較して歩く速度が遅い，朝食を抜くことが週に 3 回以上ある，現在タバコを習慣的に吸っている，お酒を飲む頻度，睡眠で休養が十分とれている

表 3: レセプトデータと特定健診データの融合方法と観測範囲

Rule to combine data		# of observation (participants)	Coverage
R	Use R only	2,415,496	100%
$R \bowtie K$	Inner join	170,240	7%
$R \rtimes\!\bowtie K$	Left outer join	2,415,496	100%
$R \bowtie \hat{K}$	Data fusion	2,415,496	100%

3.2 特定健診データ

特定健診のデータは，住民 vPID について，性別，年齢，体重，身長，血圧，血糖値，などの健康診断で測定される数値データと，飲酒の頻度などの問診結果，メタボ診断の判定結果のような医師の所見が記録されている．

本研究はこれら情報から量的変数またはカテゴリ変数として利用できる 85 項目を選び利用する．項目の一部を表 2 に示す．健診の項目は，体重や血圧のような数値データと，問診での質問項目に対する回答（はい，又はいいえの 2 択など）のカテゴリ変数からなり，これら変数を 85 次元のベクトルとして $\boldsymbol{k}$ を定義し，その集合を特定健診データ群 K と定義する．

3.3 医療需要の予測

本稿は，教師あり学習の枠組みで医療需要を予測する．ここで医療需要もレセプトのベクトルとして表現する．つまり医療需要の予測問題を，過去のレセプトの履歴 $\boldsymbol{r}$ を入力として，翌年 1 年間の発生するであろうレセプト $\boldsymbol{r}'$ を出力する予測器 $\boldsymbol{r}' = f(\boldsymbol{r})$ を構築する問題として解く．これはつまりある住民について需要を $\boldsymbol{r}'$ として予測し，その住民が実際に医療サービスを利用すればレセプトデータという観測が得られることを意味する．

ここで $\boldsymbol{r}$ の 1562 次元の要素は前述の定義から 0 または 1 の 2 値をとり，またそれぞれの要素間に依存関係はないと考えると，予測器 $f(\boldsymbol{r})$ の実体は 1562 個の 2 クラス識別器の集合となる．つまり先行研究同様に教師あり学習の方法で予測器が構築できる．

表 4: 予測器のトレーニング・テスト用データセット

	Rule to combine data		
Dataset name	Training data	Test data	Predictor
RR	R	R	$f_R(\boldsymbol{r})$
II	$R \bowtie K$	$R \bowtie K$	$f_I(\boldsymbol{r}, \boldsymbol{k})$
LL	$R \rtimes\!\bowtie K$	$R \rtimes\!\bowtie K$	$f_L(\boldsymbol{r}, \boldsymbol{k})$
LF	$R \rtimes\!\bowtie K$	$R \bowtie \hat{K}$	$f_L(\boldsymbol{r}, \hat{\boldsymbol{k}}), \hat{\boldsymbol{k}} = g(\boldsymbol{r})$
FF	$R \bowtie \hat{K}$	$R \bowtie \hat{K}$	$f_F(\boldsymbol{r}, \hat{\boldsymbol{k}}), \hat{\boldsymbol{k}} = g(\boldsymbol{r})$

3.4 異種医療データの融合

以上の定式化に従うと，本校における議論は 1562 個の 2 クラス識別器に与えるデータセットを，二つのデータ群 R，K からどうやって生成するか，という問題に帰着できる．そこで生成可能なデータセットとして表 3 の 4 種類のデータセットを考える．

一つ目のデータセットはベースラインであり，R のみを用いる．データの詳細は後述するが，この場合，利用できる情報は 2,415,496 名分のレセプトデータのみとなり，特定健診 K の情報は含まない．

残りの三つのデータセットは R と K を vPID をキーにして結合したものとなる．$R \bowtie K$ は結合の方法として内部結合を用いたデータセットである．内部結合により得られるデータセットは，レセプトと特定健診双方のドメインの情報を欠損なく有する．しかし一方で，特定健診の受診率が低いため，得られるデータは住民の 7%程度しかカバーしない．つまり患者の 93% の観測が欠落している．これは教師あり学習において大きな支障になる．

$R \rtimes\!\bowtie K$ は結合の方法として左外部結合を用いた，つまり特定健診側に欠損を許容したデータセットである．しかしこれも 93%の患者で特定健診に欠損値をもつデータセットとなり，上記の内部結合と同等の問題をかかえる可能性がある．

そこで本研究は，レセプトデータから特定健診の結果を推定する新たな推定器 $g(\boldsymbol{r})$ を考え，推定した特定健診の結果 $\hat{K}$ とレセプトデータを内部結合する $R \bowtie \hat{K}$ を提案する．観測という点では，得られるデータは $g(\boldsymbol{r})$ を介して特定健診を仮想的に観測して得られたデータと解釈できる．ここで $g(\boldsymbol{r})$ の推定性能が十分に高ければ，K と $\hat{K}$ の差は無視できる．つまり内部結合と同等の情報を得ながら左外部結合と同等の観測範囲を得たことになる．本研究はこの方法を融合と呼び，融合により仮想的に学習データを増加させることで，教師あり学習の予測性能の向上を試みる．

4 実　　験

教師あり機械学習で必要となる学習用データセットとテスト用データセットを，表 3 に示した組み合わせで生成することを考えると，可能な組み合わせは表 4 の 5 通りになる．

表 4 に示すデータセット RR はデータ融合を行わないベースラインである．RR を用いて学習した結果得られる予測器 $f_R(\boldsymbol{r})$

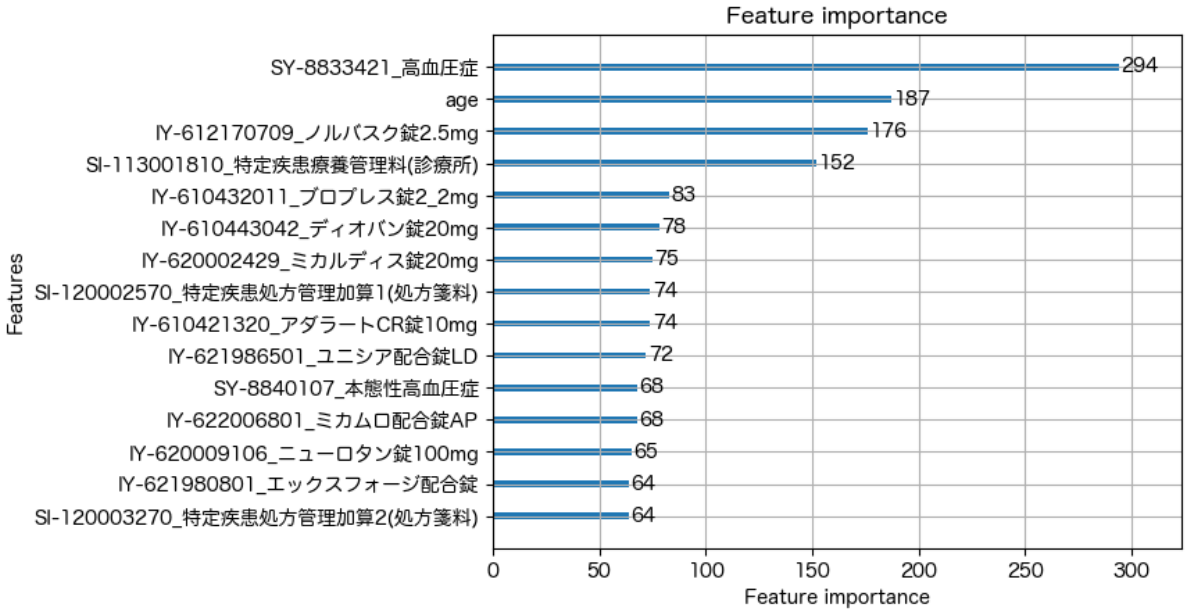


図 1: 「特定健診：服薬の有無 (血圧)」の予測に寄与したレセプト側の特徴量 (上位 15 位のみ)

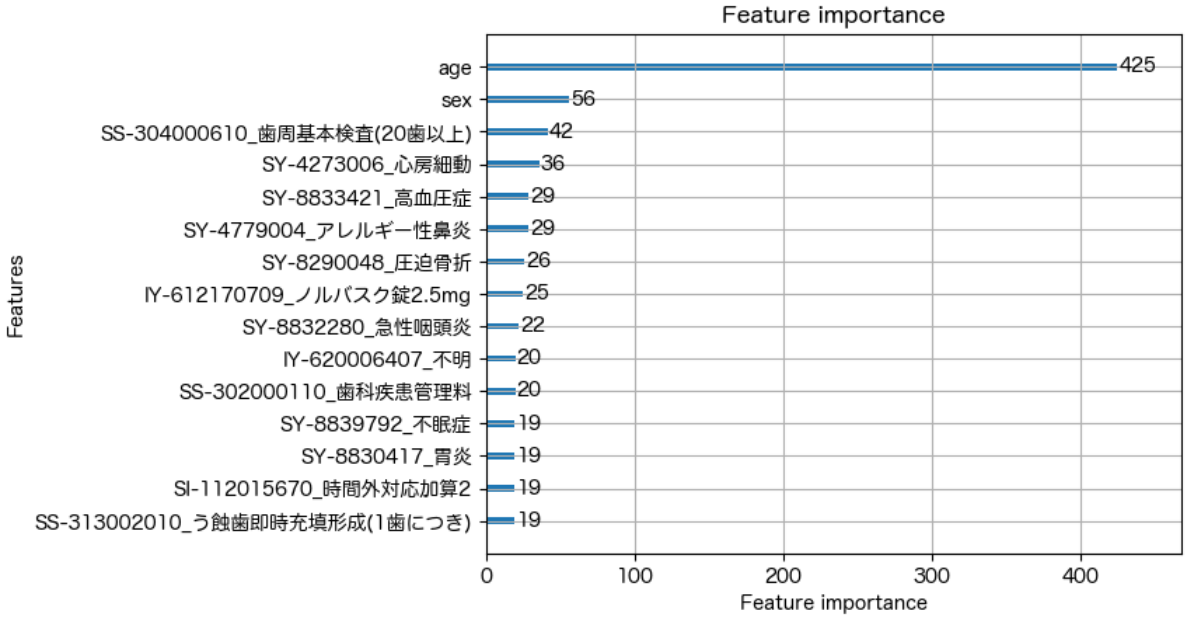


図 2: 「特定健診：身長」の予測に寄与したレセプト側の特徴量 (上位 15 位のみ)

は，過去のレセプトの履歴 $\boldsymbol{r}$ から，翌年の医療需要 $\boldsymbol{r}'$ を予測する処理となる．同様にデータセットとそれに応じて生成される予測器の対応関係は表 4 に示す通りである．

ここで，データセット LF と FF の違いが重要である．LF は学習データとして欠損を含む特定健診の結果を用いるのに対して，FF はレセプトから予測された特定健診のデータを参照するのみで特定健診のデータを直接は利用しない．予測器の構築は未知のデータに対する汎化性能を高める処理であり，その際の汎化誤差の定義が重要である．ここでデータセット LF を用いる予測器は特定健診の空間で汎化誤差を評価し，データセット FF を用いる予測器は $f_R(\boldsymbol{r})$ の出力の空間で，つまり仮想的な特定健診の空間で汎化誤差を評価する．その差を分析，議論するために本実験はデータセット LF と FF を用意している．

実験は Python 3.11 を用い，２クラス識別器は LightGBM 3.9.9 [10] を用いて実装した．使用した計算機のスペックは Intel(R) Xeon(R) CPU E7-4850 v4 （128 コア，2.10GHz），メモリ 3TB である．

教師あり学習では，全データの 80%を学習データ，残りをテストデータへとランダムに分割し使用した．予測器の学習は，評価関数を AUC（Area under the ROC curve) とし，k-分割交差検証 (k=5) を使用した．その際のパイパーパラメータの最適化は optuna [23] を用いた．予測器の性能解析には shap [24, 25] を用いた．

4.1 データセット

実験は岐阜県の保険者から提供を受けたレセプトデータ（医科，DPC，調剤，歯科），及び特定健診データを用いた．データの期間は 2014 年 4 月から 2021 年 3 月までの 7 年分であり，レセプトはのべ 2,415,496 人分，特定健診は 1,374,368 件のデータであった．

データは厚生労働省の定めるガイドラインに従って匿名化されており，本研究は，個人識別子 ID1，ID2 を手がかりに vPID を生成することで，住民の一人一人のレセプトデータと特定健診データを突合できるように前処理を行い，データセットとした．

4.2 レセプトから特定健診の推定

まずデータセット II を用いてレセプト $\boldsymbol{r}$ から特定健診の結果 $\hat{\boldsymbol{k}}$ を推定する推定器 $g(\boldsymbol{r})$ を構築した．前述のように 85 次元のベクトルである $\boldsymbol{k}$ の要素は，身長や体重などの量的変数，喫煙の有無のようなカテゴリ変数から構成されている．本実験では，前者は決定木ベースの回帰モデルで，後者は決定木ベースの多クラス識別器で推定を行う．つまり要素ごとに合計 85 個の推定器を独立に構築し，それら推定器の集合を推定器 $g(\boldsymbol{r})$ とした．

推定性能は以下の通りである．まず特定健診の項目「服薬の有無 (血圧)」,「服薬の有無 (血糖)」,「服薬の有無 (脂質)」などの項目は AUC が 0.9 以上となり高い精度で推定が可能であった．その理由は，これら項目は，レセプトに含まれる投薬履歴や医療サービスの利用履歴と直接的な関連をもつためであると考えられる．例として，特定健診側の変数「服薬の有無 (血圧)」の推定に寄与したレセプト側の変数を図 1 に示す．この図から，1) 「高血圧症」や「本態性高血圧症」の診断がついている場合，2) 高齢である場合，3)「ノルバスク錠」や「ユニシア配合錠」などの薬が処方されている場合，にその住民は「服薬の有無 (血圧)」が有となる可能性が高いことが読み取れる．ノルバスク錠やユニシア配合錠は高血圧症や狭心症に対してよく処方される薬品であることから，得られた結果は妥当であると推察される．

一方で，例えば特定健診における「身長」の推定性能は決定係数 R^2 で 0.64 程度であり，推定精度は低かった．2 に身長の推定に寄与したレセプト側の変数を示す．推定に寄与した変数としては，年齢と性別が挙げられた．これはそもそもレセプトの記録と身長には因果関係が乏しく，個人ごとの推定が困難であるためである．なぜなら，まず平均身長が世代間で大きく異なり若い世代ほど平均身長が高く，また女性よりも男性の方が平均身長が高い傾向がある．結果，年齢と性別が分かれば平均的な身長が推定できる．しかしそれ以上の個人差を推定する情報はレセプトからは得られない．そのため推定に寄与する変数は年齢と性別だけになる．

この「身長」の例は，見方を変えれば $g(\boldsymbol{r})$ には最悪の場合でも平均値で欠損値を内挿する程度の効果が期待できると言える．加えて，本来の医療需要予測の問題に対しては $g(\boldsymbol{r})$ の出力の空間で汎化誤差を小さくすることで最終的な医療需要の予測性能が向上する可能性が否定できない．そこで本稿ではこれら 85 個の推定器をそのまま $g(\boldsymbol{r})$ として利用するものとし，次に述べる医療需要の性能評価を行なった．

表 5: 性能向上した予測器の個数

	Dataset name		
	LL	LF	FF
# of predictors	350 (22%)	812 (52%)	1035 (66%)

4.3 医療需要の予測性能

図 3，表 5 に需要予測に関する性能評価の結果を示す．図 3 の左の散布図は，各点が 1562 個ある予測器の性能を表している．その X 座標はデータセット RR を用いた場合の予測性能 (AUC の値)，Y 座標はデータセットを LL に変えた場合の予測性能である．つまり性能が変化しない予測器は $y = x$ の直線上にプロットされ，データセットの変更により性能が向上した予測器は $y > x$ の領域にプロットされる．グラフは $y > x$ の領域にプロットされた予測器，つまり性能が向上した予測器をオレンジ色で，そうでない予測器を青色で可視化している．

なお提案手法を実際に利用する場合は，交差検証で一度予測性能の変化を確認し，予測性能が向上する場合にのみ提案手法を採用する必要がある．つまり最終的な予測性能はオレンジの点が示すように予測性能が向上する場合と，$y = x$ 直線上の点，つまり提案手法を使わずに予測性能も変化しない場合とのどちらかになるようにする．厳密には過学習により性能が低下する可能性は残されるが，過学習の問題は教師あり学習を用いた手法全般に関する問題であり，学習時のハイパラーパラメータの適切な調整と，適切な手順での交差検証により十分に回避できると考えられる．以上の点を踏まえた上で，本節では性能が下がる場合上がる場合双方を分析した．

表 5 は，性能向上した予測器の個数，つまりグラフにおけるオレンジの点の数をまとめたものである．特定健診の欠損を含むデータセットで学習した予測器 f_L で，欠損を含むデータセット LL を予測した結果は，性能が向上した予測器は 350 個，全体の 22%にとどまった．またその性能向上率も僅かであり，ほぼベースラインの RR と同等である．一方で図 3 の左図の青色の点群が示すように性能が低下した予測器も多数見受けられた．これはデータの欠損の影響だと考えられる．

図 3 の中央は，予測器 f_L は変更せずに，$g(\boldsymbol{r})$ の推定結果を用いて予測器への入力を $\boldsymbol{k}$ から $\hat{\boldsymbol{k}}$ へ変更したケースである．青色の点の分散が広がって見えるのは，学習時に用いたデータ群 K と予測時に用いるデータ群 $\hat{K}$ に乖離があるため，予測器によっては AUC の値の変動が大きくなり，y 軸方向の分散が大きくなるためと考えられる．定量的には表 5 を見ると 812 個 (52%) の識別器で性能が向上しており，$\hat{\boldsymbol{k}}$ の利用が性能向上に寄与しうることを示唆している．

図 3 の右は，さらに予測器を f_L から f_F に変更したケースである．図中央と右図を比較すると，右図の方が点群の分散が小さくなっているのは，予測器を $\hat{K}$ の空間で学習し，$\hat{K}$ の空

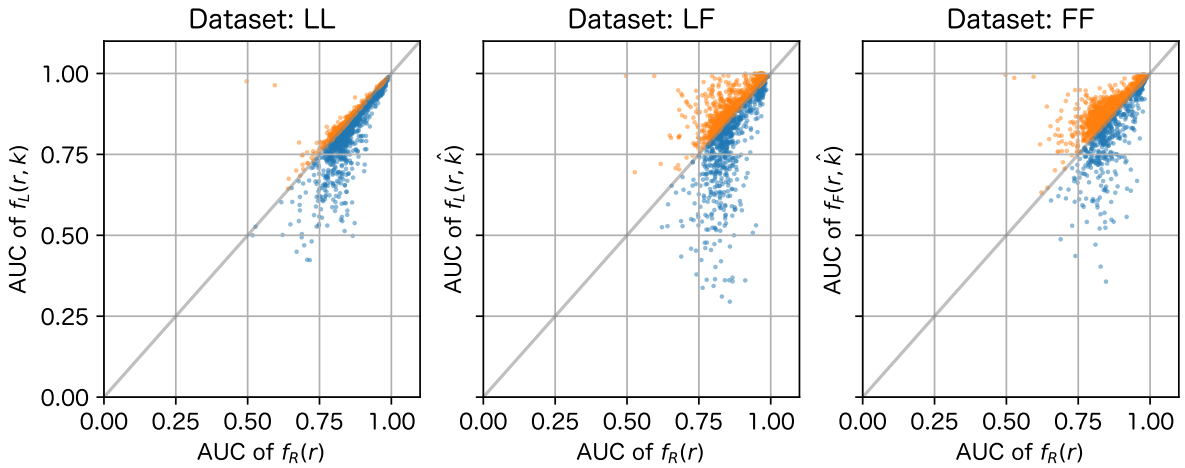


図 3: 予測器の性能評価

間で使用するため，予測器間での AUC の値の変動は小さくなるためである．また全体的に性能が向上しており，定量的には表 5 で見ると 1035 個 (66%) の予測器で性能が向上する結果となっている．

全体で見ると表 5 が示すように，データセット FF を用いる場合が最も良い予測性能が得られた．これはつまり，レセプトから特定健診の結果を予測し，予測された特定健診の結果とレセプトデータをくみわせる方法が，最も医療需要予測に適していたことになる．

4.4 予測性能に対するデータ融合の貢献

次に，なぜ予測性能が変化したか，その原因の分析を行った．図 4 は，データ融合の有無により，予測に寄与する特徴量がどのよう変化するか示した図である．

予測する医療需要は「初期加算（リハビリテーション料)」である．これは医療機関がリハビリの指導を行うと請求されるサービスであり，リハビリ指導の需要を反映している変数である．ここでデータ融合を用いない予測器 f_R と，データ融合を用いる予測器 f_F で予測に寄与した特徴量を比較すると，二つの予測器は全く異なる特徴量を手がかりとして予測を行なっていることが明らかになった．

図 4a は，データ融合を用いない予測器 f_R で寄与に寄与した上位 15 個の特徴量である，この図から，年齢，そしてパーキンソン病や腰痛症のような比較的高齢者に多い傷病が予測に利用されていることがわかる．

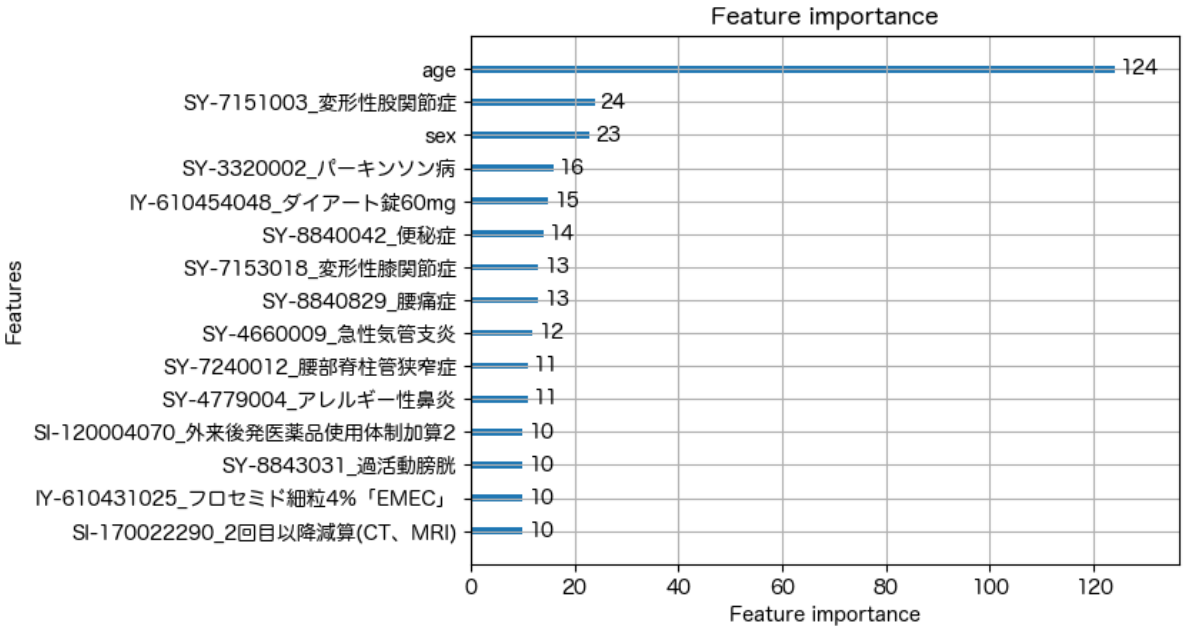
一方で，図 4b のデータ融合を用いる予測器は，手がかりとして $g(\boldsymbol{r})$ で推定した特定健診の結果から「歩行速度」や「歩行または身体活動」の項目を参照している．これは日常生活における歩行速度や身体活動の問題を尋ねる項目であり，リハビリテーションという医療サービスへの需要を直接聞く質問項目になっている．二つの予測器の性能を比較すると前者は AUC で 0.76 であり，データ融合を用いる後者は 0.86 と大幅な性能向上が見られる．詳細を見ると，データ融合を用いない予測器はパーキンソン病や変形性股関節症の有無を予測に利用している．しかし，パーキンソン病や変形性股関節症の診断を受ける患者は全体で見るとごく一部である．つまり特異な事例に重点を置いて，再現性 (Recall) が高くなるようにデータを学習している可能性が考えられる．一方で，データ融合を用いた予測器は，特定健診の結果「歩行速度」と「歩行又は身体活動」の項目を参照している．これは特定健診の問診で得られる項目であり，日常生活で患者が歩行速度や身体活動に困難さを感じていることを表す項目である．これはまさに歩行に関するリハビリ指導への需要であり，その需要の有無が間接的にレセプトから予測されて，最終的な需要予測に利用されていることが推察される．

5 考　　察

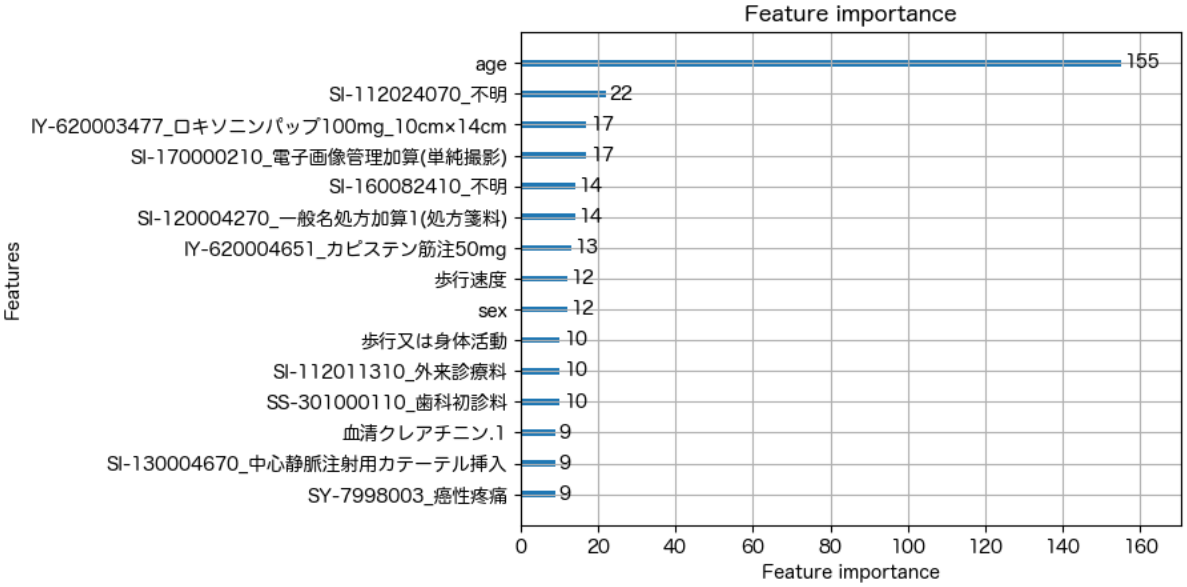
教師あり学習の手法として見ると，我々が提案したデータ融合手法は交差検証により性能向上が認められる場合のみに採用すれば良い．つまり表 5 で性能向上が認められた 66%の識別器でのみ提案手法を採用し，残りの 34%は従来法を用いれば，全ての予測においてデメリットを生じずに提案手法の恩恵が得られる．この点で提案法は医療情報の予測を扱う際の実用的手法になると期待できる．

次に機械学習の手法として提案法により予測精度が向上した理由を考察する．まず提案手法はデータ拡張 (Data augmentation) や転移学習の一手法であると解釈することができる．機械学習におけるデータ拡張とは学習データ数を人為的に水増しすることで汎化性能を向上させる手法である．具体的には画像認識における AlexNet [26] などが有名である．画像処理におけるデータ拡張は，学習データの画像を左右反転や回転させたり輝度を変化させるなど，画像から画像への単純な変換処理の組み合わせとして実現されている．一方で，提案手法は $g(\boldsymbol{r})$ を用いることで実際には存在していない特定健診の結果を推定しその推定結果で学習データを増やしている．これはデータを仮想的に水増しするという点でデータ拡張の一種であると言える．

データは患者を観測して得られた観測結果と捉えると，レセプトデータはレセプトドメインで患者を観測して得られたデータ，特定健診は特定健診のドメインで患者を観測して得られた



(a) 予測器 f_R に寄与した特徴量（上位 15 位のみ）



(b) 予測器 f_L に寄与した特徴量（上位 15 位のみ）

図 4: 「初期加算（リハビリテーション料)」の予測に寄与した特徴量

データである．転移学習の文脈で考えると，$g(\boldsymbol{r})$ には特定健診ドメインの知識が含まれており，$g(\boldsymbol{r})$ を介してレセプトデータ上の患者を再観測した結果レセプトデータの解釈が変わり予測精度が向上した，と解釈することもできる．この点は，今後はさらなる分析を行い定量的・理論的背景を明らかにすることが必須であり，これは本研究の今後の課題の一つである．

医療データのスパース性に関しては，従来研究ではデータの時系列をより広く観測し統合することで個人追跡を行い，精緻な名寄せにより利用可能なデータ数を増加させる方法が取られている [22]．一方で本研究での提案法は，仮想的な観測を加えることで利用可能なデータ数を増加させている．提案法と従来法は相補的な関係にあり，両者は併用・共存することで，スパース性に起因した課題がより緩和できると期待できる．

6 おわりに

本稿は，医療データから将来の医療需要を個人単位で予測する問題を考え，その予測精度を向上させる方法としてレセプトと特定健診という二つの異種医療データを融合する方法を提案した．実験では，7 年間にわたる約 241 万人のデータを用いて 1562 項目 の医療需要を個人単位で予測し，結果，提案手法により 1035 項目 (66%) で予測精度が向上することが示された．また予測精度が向上するメカニズムを分析し，医療需要予測のための最適なデータ融合方法について考察を行った．

本研究の今後の課題としては，精度向上のメカニズムの詳細な分析があげられる．考察で述べたように，そのメカニズムは機械学習における転移学習やデータ拡張と関連していることが予想される．また一般化して考えると深層学習における Encoder モデル [27] や Attension モデル [28] を用いることで本手法と同等の結果が得られる可能性もある．この点は本研究の今後の課題としたい．

一方で提案法の有用性は既に実験で実証されており，本提案手法は医療情報を扱う際の要素技術になると我々は期待している．今後提案手法の応用例として，大規模かつ精緻な医療需要予測にも取り組む予定である．

文　献

[1] S. Matsuda and K. Fujimori: “The claim database in japan”, Asian Pacific Journal of Disease Management, **6**, 3–4, pp. 55–59 (2014).

[2] 満武：“ウェルネスのための ICT：2. 日本のレセプト情報・特定健診等データベース（NDB）の有効活用”, 情報処理, **56**, 02, pp. 140–144 (2015).

[3] E. W. Gregg, J. P. Boyle, T. J. Thompson, L. E. Barker, A. L. Albright and D. F. Williamson: “Modeling the impact of prevention policies on future diabetes prevalence in the united states: 2010–2030”, Population health metrics, **11**, 1, pp. 1–9 (2013).

[4] T. M. Dall, M. V. Storm, R. Chakrabarti, O. Drogan, C. M. Keran, P. D. Donofrio, V. W. Henderson, H. J. Kaminski, J. C. Stevens and T. R. Vidic: “Supply and demand analysis of the current and future us neurology workforce”, Neurology, **81**, 5, pp. 470–478 (2013).

[5] J. Spetz, L. Trupin, T. Bates and J. M. Coffman: “Future demand for long-term care workers will be influenced by demographic and utilization changes”, Health Affairs, **34**, 6, pp. 936–945 (2015).

[6] B. Klute, A. Homb, W. Chen and A. Stelpflug: “Predicting outpatient appointment demand using machine learning and traditional methods”, Journal of medical systems, **43**, pp. 1–10 (2019).

[7] A. Earnest, M. I. Chen, D. Ng and L. Y. Sin: “Using autoregressive integrated moving average (arima) models to predict and monitor the number of beds occupied during a sars outbreak in a tertiary hospital in singapore”, BMC Health Services Research, **5**, 1, pp. 1–8 (2005).

[8] Y. Huang, C. Xu, M. Ji, W. Xiang and D. He: “Medical service demand forecasting using a hybrid model based on arima and self-adaptive filtering method”, BMC Medical Informatics and Decision Making, **20**, 1, pp. 1–14 (2020).

[9] J. H. Friedman: “Greedy function approximation: a gradient boosting machine”, Annals of statistics, pp. 1189–1232 (2001).

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu: “Lightgbm: A highly efficient gradient boosting decision tree”, Advances in neural information processing systems, **30**, pp. 3146–3154 (2017).

[11] V. Kecman: “Support vector machines – an introduction”, Support vector machines: theory and applications, Springer, pp. 1–47 (2005).

[12] W. Dai, T. S. Brisimi, W. G. Adams, T. Mela, V. Saligrama and I. C. Paschalidis: “Prediction of hospitalization due to heart diseases by supervised learning methods”, International journal of medical informatics, **84**, 3, pp. 189–197 (2015).

[13] J. Sato, N. Mitsutake, M. Kitsuregawa, T. Ishikawa and K. Goda: “Predicting demand for long-term care using japanese healthcare insurance claims data”, Environmental Health and Preventive Medicine, **27**, pp. 42–42 (2022).

[14] S. Kaushik, A. Choudhury, P. K. Sheron, N. Dasgupta, S. Natarajan, L. A. Pickett and V. Dutt: “AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures”, Frontiers in big data, **3**, p. 4 (2020).

[15] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie and L. Jorm: “Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk”, Scientific reports, **10**, 1, p. 1111 (2020).

[16] A. Zeroual, F. Harrou, A. Dairi and Y. Sun: “Deep learning methods for forecasting COVID–19 time-series data: A comparative study”, Chaos, Solitons & Fractals, **140**, p. 110121 (2020).

[17] F. Finger, S. Funk, K. White, M. R. Siddiqui, W. J. Edmunds and A. J. Kucharski: “Real–time analysis of the diphtheria outbreak in forcibly displaced myanmar nationals in bangladesh”, BMC medicine, **17**, 1, pp. 1–11 (2019).

[18] M. J. Keeling, L. Dyson, G. Guyver-Fletcher, A. Holmes, M. G. Semple, I. Investigators, M. J. Tildesley and E. M. Hill: “Fitting to the UK COVID-19 outbreak, short-term forecasts and estimating the reproductive number”, Statistical Methods in Medical Research, **31**, 9, pp. 1716–1737 (2022).

[19] M. W. Gardner and S. Dorling: “Artificial neural networks (the multilayer perceptron) — a review of applications in the atmospheric sciences”, Atmospheric environment, **32**, 14-15, pp. 2627–2636 (1998).

[20] F. Bagattini, I. Karlsson, J. Rebane and P. Papapetrou: “A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records”, BMC medical informatics and decision making, **19**, pp. 1–20 (2019).

[21] 佐藤, 山田, 合田, 喜連川, 満武：“保険医療データに於ける複数個人を包含する暗号化された識別子の検出方法の検討”, 電子情報通信学会データ工学研究会，電子情報通信学会技術報告, **27**, 11 (2018).

[22] J. Sato, H. Yamada, K. Goda, M. Kitsuregawa and N. Mitsutake: “Enabling patient traceability using anonymized personal identifiers in japanese universal health insurance claims database”, AMIA Joint Summits on Translational Science proceedings, pp. 345–352 (2019).

[23] T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama: “Optuna: A next-generation hyperparameter optimization framework”, Proc. of International Conference on Knowledge Discovery and Data Mining (SIGKDD’19) (2019).

[24] S. M. Lundberg and S.-I. Lee: “A unified approach to interpreting model predictions”, Proc. of International Conference on Neural Informatics Processing Systems (NIPS’17), Curran Associates, Inc., pp. 4765–4774 (2017).

[25] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee: “From local explanations to global understanding with explainable AI for trees”, Nature Machine Intelligence, **2**, 1, pp. 2522–5839 (2020).

[26] A. Krizhevsky, I. Sutskever and G. E. Hinton: “Imagenet classification with deep convolutional neural networks”, Communications of the ACM, **60**, 6, pp. 84–90 (2017).

[27] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova: “BERT: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805 (2018).

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin: “Attention is all you need”, Advances in neural information processing systems, **30**, (2017).