

医療費データと歩数データの結合分析に関する検証： 共通属性とモデルベース統合を用いた部分データ合成

渋谷 亮^{1,a)} 波多野 卓磨^{2,b)} 佐々木 誓良³ 小野里 尚紀³ 藤澤 陽介³ 大関 晃一² 下出 隼大²
小藺井 智之²

概要：本研究では「医療費の情報を含むデータセット A」と「保険会社で保有する歩数の情報を含むデータセット B」を、両者共通の健康診断項目で組み合わせ、医療費と運動習慣の関連を解析した。結合には共通属性に基づく2つのモデルベース統合を用いた。マクロアプローチでは層別セルごとにAの医療費平均とBの歩数平均を対応付けた。ミクロアプローチではAで学習した回帰モデルをBに適用することでBに医療費列を合成し、擬似的に医療費と歩数を同時に観測したデータを構築した。両手法による分析結果を比較して医療費と歩数の関係を検討した。補足として回帰モデルに差分プライバシーを適用し、追加的なプライバシー強化が推定に与える影響を評価した。

キーワード：部分データ合成, 差分プライバシー, プライバシー保護機械学習, 統計的マッチング

Evaluation of Joint Analysis of Medical Expenditure and Step Count: Partially Synthetic Data via Model-Based Integration Using Common Attributes

RYO SHIBUYA^{1,a)} TAKUMA HATANO^{2,b)} CHIKARA SASAKI³ NAOKI ONOZATO³ YOSUKE FUJISAWA³
KOICHI OZEKI² SHUNTA SHIMODE² TOMOYUKI OSONOI²

Received: xx xx, xxxx, Accepted: xx xx, xxxx

Abstract: This study analyzes the association between medical expenditures and physical activity by integrating two datasets: Dataset A, containing medical expenditure information, and Dataset B, containing step count data collected by an insurance company. Integration was performed using common attributes shared by both datasets through two model-based approaches. In the macro approach, the mean medical expenditure from Dataset A was paired with the mean step count from Dataset B for each stratified cell. In the micro approach, a regression model was trained on Dataset A and applied to Dataset B to impute a medical expenditure variable, creating a pseudo dataset with simultaneous observations of medical expenditures and step counts. Results from both approaches were compared to investigate the relationship between medical expenditures and step counts. Additionally, differential privacy was applied to the regression model to assess the impact of enhanced privacy protection on estimation outcomes.

Keywords: Partially Synthetic Data, Differential Privacy, Privacy Preserving Machine Learning, Statistical Matching

¹ スミセイ情報システム株式会社
Sumitomo Life Information Systems Co.,Ltd.
² 日鉄ソリューションズ株式会社
NS Solutions Corporation
³ 住友生命保険相互会社
Sumitomo Life Insurance Company
a) riyoushibuya@am.sumitomolife.co.jp
b) hatano.takuma.hq2@jp.nssol.nipponsteel.com

1. はじめに

複数機関のヘルスケアデータを統合して分析する研究では、単一データでは捉えにくい生活習慣と医療アウトカムの関係が報告されている。代表例として、UK Biobank[1]

では手首装着型加速度計による活動量データを NHS の入院記録 (HES) に連結し、主要な入院理由 (25 項目) に対するリスクを解析している。また、米国の NIH All of Us[2] では、医療機関の電子カルテ (EHR) と Fitbit 由来のウェアラブルデータを結合し、歩数や心拍などの指標と診療記録を縦断的に評価する研究が進んでいる。こうした異機関データ連携・異種データ統合の進展は、個々のリスクプロファイルに応じた予防介入や保険商品設計といったデータドリブンなヘルスケア施策の創出を後押しすると期待される。

一方、異機関データの統合では共通 ID がなく、レコードリンクに起因する未結合や誤結合が推定に偏りを生じさせ得る。これを回避する設計としては、集計指標を作成して照合する集計レベルの統合 (マクロアプローチ) と、個票に未観測量を合成して擬似的な同時観測を得る個票レベルの統合 (ミクロアプローチ) がある [3]。前者は実装が簡便な反面、層内の変動や共分散構造が失われ、条件付きの関係性や効果の推定が難しくなる。後者は、共通の説明変数を手掛かりに未観測量を推定・補完することで個票の擬似的同時性を確保でき、層別や共変量調整を伴う解析に適している。

本研究では、データセット A (健診共通項目に加えて年間医療費を含む) とデータセット B (健診共通項目に加えてウェアラブル由来の歩数を含む) の結合分析を行う。両データに共通 ID は存在せず、同一個人が含まれるかは不明である。結合にあたっては、共通属性を用いて先述の 2 つの統合手法を並行して適用する。(1) マクロアプローチでは、共通属性により層別セルを構成し、各セルで A の医療費平均と B の歩数平均を算出して照合する。(2) ミクロアプローチでは、今回は A で共通属性から医療費を予測するモデルを学習し、B の共通属性を予測モデルに適用することで、擬似的に医療費と歩数の同時観測データを得る。(1)(2) の両手法から得られる結果を比較し、平均歩数と医療費の関係を議論する。

ミクロアプローチにおける回帰モデルの学習・適用に際しては、モデルの個人情報該当性に配慮する。本研究で用いるデータセット A は匿名加工情報として適切に加工されたものを利用しており、実験はデータ提供元との協議・承認のもとで実施している。作成した回帰モデルは公開せず、データセット B への適用と集計指標の作成に限定して利用する。医療費予測に関する既存研究 [4] を参考にして採用した回帰モデルである勾配ブースティング木 (XGBoost) [5] は分割閾値と葉の統計量から構成され、個票の値を直接保持しないため、これらの運用と併せて個票の再構成リスクは実務上十分に低減され则认为する。今回は、補足的な分析として、学習時の勾配クリッピングとノイズ付与を用いる差分プライバシー版 XGBoost[6] を適用し、複数のプライバシー水準 ϵ で医療費の予測誤差および連携分析結

果の変化を評価する。得られた結果は、追加的なプライバシー強化が推定結果に与える影響を定量的に把握するための参考結果として位置づける。

2. 実験手法

2.1-2.3 にて、用いたデータセットの詳細を説明する。2.4 以降にて、統合分析を行う手法について説明する。

2.1 データセット A (共通属性+年間医療費) の詳細

本研究ではデータセット A として、株式会社 JMDC が保有する「JMDC Claims Database」(以下 JMDC データベース)[7] を使用する。JMDC データベースの特徴は 2 点ある。1 つ目は JMDC 社独自の医療辞書により、傷病、薬剤、診療行為情報が自動標準化されていること。2 つ目は個人を特定する情報は不可逆的匿名化技術による高度暗号化処理済であること。転院や複数機関の受診があっても追跡可能であり、民間利用可能な日本最大規模の疫学レセプトデータベースとなっている。

本研究に用いた JMDC データベースは、複数の健康保険組合から収集された匿名化処理が施された医療ビッグデータであり、健康保険組合に所属する加入者 (従業員およびその被扶養者) を対象とする。収録されている内容は、加入者が医療機関を受診した際に発行される全レセプトデータ (医科レセプトおよび調剤レセプト) に加え、健保組合が保有する全健康診断結果 (特定健診や健保組合が実施する定期健診の結果等を含む) のうち健保組合から JMDC へ連携された健診結果を含んでおり、医療受診情報 (診療日、診療科、処方薬剤、入院・外来区分など) とともに、健診における各種検査値 (BMI、腹囲、血圧、ヘモグロビン A1c (HbA1c)、血清脂質など) および生活習慣に関する問診項目 (喫煙、飲酒、運動習慣など) が網羅されている。

データセット A は 2023 年 10 月から 2024 年 9 月に在籍している人を対象としており、機械学習実行時には 2022 年 10 月から 2023 年 9 月に在籍していること、2022 年 10 月から 2023 年 9 月中の入院や手術が無いことも条件とする。BMI、収縮期血圧 (SBP)、拡張期血圧 (DBP)、LDL コレステロール、ヘモグロビン A1c (HbA1c)、尿蛋白 (UP) のいずれかの健診項目に欠損値 (Null) がある人は除外される。また、空腹時血糖 (FBS) に関しては Null が許容され、性別と年齢、BMI の 1 刻み区分の 3 つが一致しているレコードの、空腹時血糖 (FBS) の平均値で補完される。また、ICD10 コードで「S00~T98: 損傷、中毒およびその他の外因の影響」、「O00~O99: 妊娠、分娩および産じょく (褥)」、「F00~F99: 精神および行動の障害」による医療費は除外した。

説明変数としては最新の健診受診日について BMI、収縮期血圧 (SBP)、拡張期血圧 (DBP)、HDL コレステロール、空腹時血糖 (FBS)、ヘモグロビン A1c (HbA1c)、尿

蛋白（UP）を取得し、目的変数は2023年10月から2024年9月の間の医科および調剤レセプトの点数合計値を使用する。医科および調剤レセプトの点数は、医療サービスや処方された薬剤に基づいて算定されるものであり、点数1点は10円に相当するため、最終的な医療費は点数合計値に10を掛けた値が実際の金額に相当する。

2.2 データセット B（共通属性+歩数）の詳細

本研究におけるデータセット B は、住友生命が2018年に導入した健康増進型保険「Vitality」[8]に基づく健康プログラムに参加している人を対象としている。このプログラムは、加入者が日常的に健康増進活動を行うことを促進し、その活動をポイント化することにより、健康的な行動の変容を目指している。具体的には、加入者は生活習慣のチェックや健康診断、予防検診を通じて自身の健康状態を把握し、得られたデータをもとに実施した活動に応じたポイントを獲得する。このポイントは、翌年度の保険料の割引率や特典に反映され、参加者の健康状態の改善を促す仕組みが整えられている。

データセット B は2024年1月以前に在籍している人を対象としており、同様にBMI、収縮期血圧（SBP）、拡張期血圧（DBP）、LDL コレステロール、ヘモグロビン A1c（HbA1c）、尿蛋白（UP）のいずれかの健診項目が欠損値（Null）である人は除外される。空腹時血糖（FBS）の欠損値（Null）は許容され、性別と年齢、BMI の1刻み区分ごとの空腹時血糖（FBS）の平均値で補完される。取得変数としてはBMI、収縮期血圧（SBP）、拡張期血圧（DBP）、LDL コレステロール、空腹時血糖（FBS）、ヘモグロビン A1c（HbA1c）、尿蛋白（UP）を取得し、さらに2024年1月から2024年12月中の1日当たり平均歩数を加える。歩数はウェアラブルデバイスまたはスマートフォンアプリにより計測している。取得したすべての健診値については、厚生労働省の「標準的な健診・保健指導プログラム」[9]に基づいて異常値を除外する。

2.3 データセット A と B の基本統計量

本研究の分析対象となったデータセット A（N=500,103）およびデータセット B（N=577,446）の基本統計量をそれぞれ表1、表2に示す。

表 1 データセット A 基本統計量（レコード数 = 500,103）

	BMI	拡張期 血圧	収縮期 血圧	空腹時 血糖	LDL コレステロール
平均	22.8	72.6	118.4	92.8	114.5
中央値	22.0	72.0	118.0	90.3	112.0
25%点	20.0	64.0	107.0	86.0	93.0
75%点	25.0	80.0	128.0	97.0	134.0

表 2 データセット B 基本統計量（レコード数 = 577,446）

	BMI	拡張期 血圧	収縮期 血圧	空腹時 血糖	LDL コレステロール
平均	23.5	75.5	121.9	96.2	121.6
中央値	22.9	75.0	121.0	93.3	118.0
25%点	20.7	68.0	111.0	87.0	98.0
75%点	25.5	83.0	131.0	101.0	141.0

2.4 マクロアプローチでの統合分析手法

データセット A と B の統合分析の手法を説明する。まずマクロアプローチの手法では、データセット B において以下の変数に基づいて健診値の集計結果を取得した：

- 年齢区分 [35 歳以下, 36–45 歳, 46–55 歳, 56–65 歳, 66 歳以上]
- 性別区分 [男性, 女性]
- 1 日当たり平均歩数の区分 [4000 歩以下, 4000–6000 歩, 6000–8000 歩, 8000–10000 歩, 10000–12000 歩, 12000 歩以上]

これに加え、BMI、拡張期血圧、収縮期血圧、LDL コレステロール、血糖については、各区分の平均値と標準偏差を取得し、マッチングの範囲を平均値±標準偏差とした。血糖は空腹時血糖を用い、尿蛋白は（-）を正常、（±, +, ++, +++）を異常とする二区分でマッチングを行った。これらのマッチング条件に合致するレコードをデータセット A から抽出し、医療費の平均を取得し、予測値として利用した。

2.5 ミクロアプローチでの統合分析手法

ミクロアプローチの手法では、データセット A と B の両方に含まれる年齢、性別、健診値情報（BMI、拡張期血圧、収縮期血圧、LDL コレステロール、空腹時血糖、尿蛋白）を特徴量として、データセット A を用いて医療費を予測するモデルを構築した。医療費予測モデルを構築するにあたって、先行研究 [4] を参考にした。この研究では、Two-Part モデルを用いて医療費の予測を行っている。本研究においても Two-Part モデルを採用し、以下の2段階で予測を行った：

- 第1段階：医療費が0になるか正の値になるかを予測する分類モデルをXGBoostで構築する。このモデルは、医療費の発生確率を出力する。
- 第2段階：医療費が正の値であるデータに基づいて回帰モデルをXGBoostで構築する。このモデルは、医療費の予測値を出力する。

最終的な予測値は、第1段階で得られた確率と第2段階の出力を乗じて算出した。

2.6 補足実験：差分プライバシーを用いたマイクロアプローチ

2.5 節の XGBoost を DP-XGBoost に置き換え、差分プライバシーを導入したマイクロアプローチを実施した。

差分プライバシー法のノイズ効果が第 1 段階にも第 2 段階にも加えるとノイズ効果が分離できず、結果の解釈が難しくなると判断したため、本研究では第 1 段階の分類モデルでは DP-XGBoost は採用せず、通常の XGBoost (non-DP) を適用した。そして第 2 段階では DP-XGBoost と XGBoost (non-DP) を併用し、モデル種別と ϵ の値の違いによる予測結果の変化を評価した。

2.7 ハイパーパラメータのチューニング方法

2.5 節および 2.6 節で用いた XGBoost および DP-XGBoost のハイパーパラメータは、Optuna[10] により最適化した。調整した主なパラメータは、`learning_rate`, `max_depth`, `lambda`, `alpha`, `subsample`, `min_child_weight`, `colsample_bytree` である。最終的なモデルの予測性能は RMSE (平均 2 乗誤差) を用いて評価した。

3. 実験結果

3.1 マクロアプローチでの統合分析の結果

集計レベルでデータセット A と B の統合分析を行った結果、56～65 歳区分以上の高齢区分については平均歩数が多いほど年間医療費が低い傾向が確認できた。65 歳以上区分は N 数が少ないため、本稿では男女別、56～65 歳、尿蛋白正常者の 1 日当たり平均歩数別の年間医療費の中央値 (Median) を図 1, 図 2 に示す。両図から明らかなように、平均歩数が歩数が多いほど年間医療費が低い傾向が確認された。この結果は、生活習慣の改善が医療費の削減に寄与する可能性を示唆している。

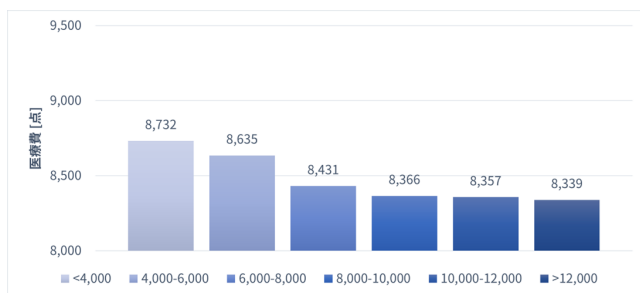


図 1 マクロアプローチにて得られた、男性 56～65 歳、尿蛋白正常者の 1 日当たり平均歩数別の年間医療費の中央値 (Median)。

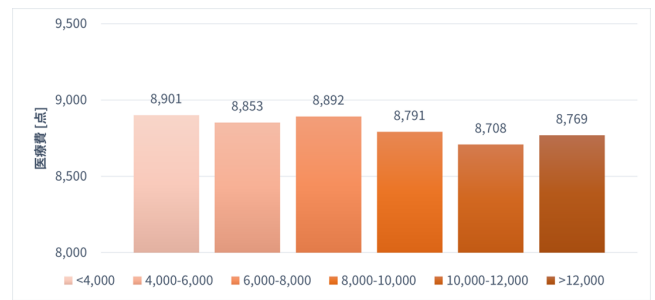


図 2 マクロアプローチにて得られた、女性 56～65 歳、尿蛋白正常者の 1 日当たり平均歩数別の年間医療費の中央値 (Median)

3.2 ミクロアプローチでの統合分析の結果

分類モデルである第 1 段階の予測精度は $AUC = 0.638$ であった。第 2 段階まで実施して統合分析を行った結果、平均歩数が多いほど年間医療費が低い傾向が明らかとなった。男女別、56～65 歳、尿蛋白正常者の 1 日当たり平均歩数別の年間医療費の中央値 (Median) を図 3, 図 4 に示す。両図から明らかなように、平均歩数が多いほど年間医療費が低い傾向が確認された。この結果は、3.1 の結果と同様、生活習慣の改善が医療費の削減に寄与する可能性を示唆している。

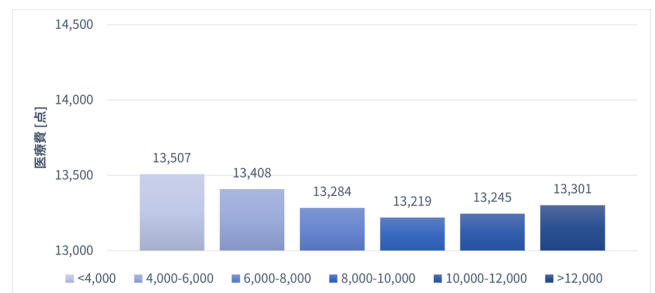


図 3 ミクロアプローチにて得られた、男性 56～65 歳、尿蛋白正常者の 1 日当たり平均歩数別の年間医療費の中央値 (Median)

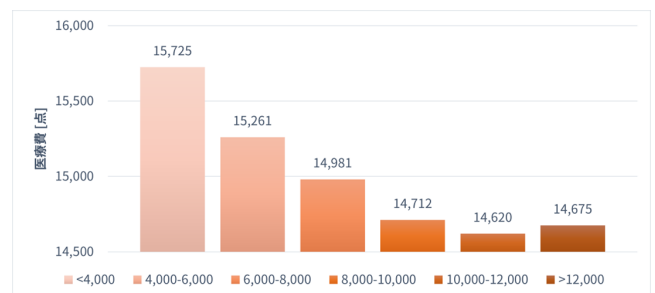


図 4 ミクロアプローチにて得られた、女性 56～65 歳、尿蛋白正常者の 1 日当たり平均歩数別の年間医療費の中央値 (Median)

3.3 ミクロアプローチの補足実験：差分プライバシー基準の適用

プライバシーパラメータ ϵ を変化させながら、DP-XGBoost によって医療費を予測した際の誤差を RMSE

にて評価した結果を図5に示す。εを大きくするとRMSEの減少が一定の水準で落ち着く傾向が確認できる。また、XGBoost (non-DP) を用いて医療費を予測した際の誤差とも漸近する様子が確認できた。Total ε = 0.0138にてRMSEが十分に小さくなったと判断し、本パラメータを用いて後続の統合分析を行う。

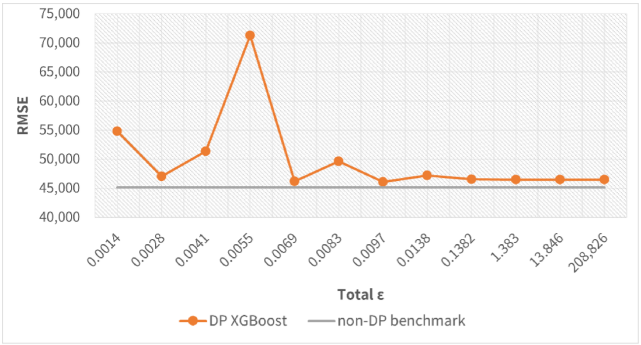


図5 プライバシー水準 ε を変化させたときの RMSE の変化

統合分析結果を図6と図7に示す。図が示すように、DP-XGBoostのプライバシー水準 Total ε = 0.0138の際に得られた実験結果についても、平均歩数が多いほど年間医療費が低い傾向が見られた。しかし、縦軸に示す医療費を予測した値を観察すると、図3と図4を比較した場合に値のオーダーが2桁程度小さいことがわかる。この点については、後掲の3.4にて詳細を確認する。

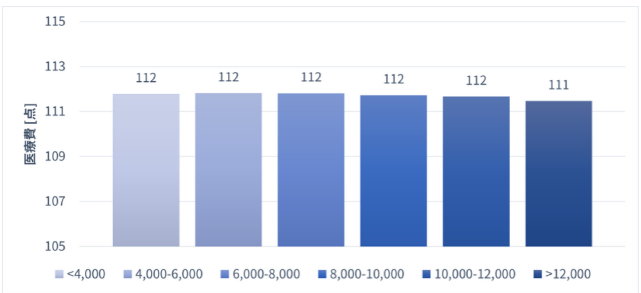


図6 ミクロアプローチ (差分プライバシー適用) にて得られた、男性 56~65 歳、尿蛋白正常者の 1 日当たり平均歩数別の年間医療費の中央値 (Median)

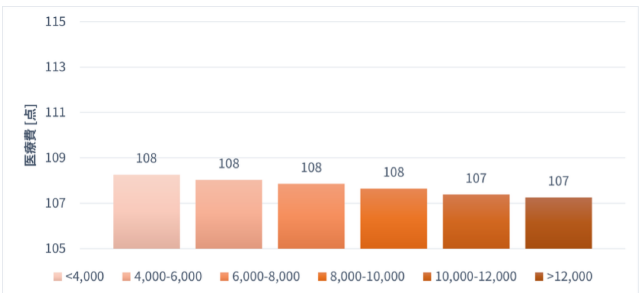


図7 ミクロアプローチ (差分プライバシー適用) にて得られた、女性 56~65 歳、尿蛋白正常者の 1 日当たり平均歩数別の年間医療費の中央値 (Median)

3.4 補足実験結果の追加調査

医療費の予測に関する追加調査を実施した。まず、データセット A における実際の年間医療費のヒストグラムを図8に示す。この分布は0付近に単一のピークを持ち、右側に長い裾を有するロングテール形状であることが確認できる。

次に、XGBoost を用いた予測年間医療費のヒストグラムを図9に示す。分布のピーク位置はデータセット A と同様に0付近にあり、形状も単峰性かつロングテールを示している。全体として、XGBoost による予測分布は実測値の特徴を概ね再現していると言える。

一方、DP-XGBoost (Total ε = 0.01384) を用いた予測年間医療費のヒストグラムを図10に示す。ピーク位置は同様に0付近であるものの、全体形状は二峰性となり、低額域の右側に第2のピークが出現している。さらに、分布全体のスケールは実測値や通常 XGBoost の結果に比べて小さくなっていることが確認できる。

以上より、DP-XGBoost の結果は、実測値や通常 XGBoost の結果と分布形状およびスケールの両面で大きく異なっている。このため、本手法による医療費予測は、結合分析の前提となる精度を十分に満たしていないと判断される。今後は、この差異が生じた要因について検証を行い、適切な予測性能を確保するための条件や手法を検討する必要がある。

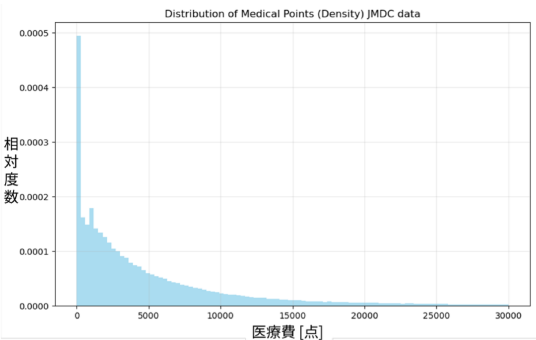


図8 データセット A における年間医療費のヒストグラム

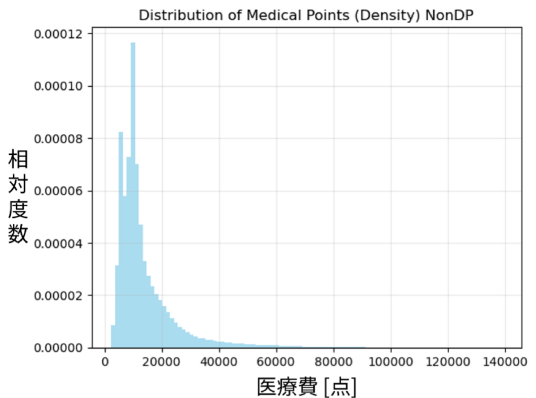


図9 XGBoost で予測した年間医療費のヒストグラム

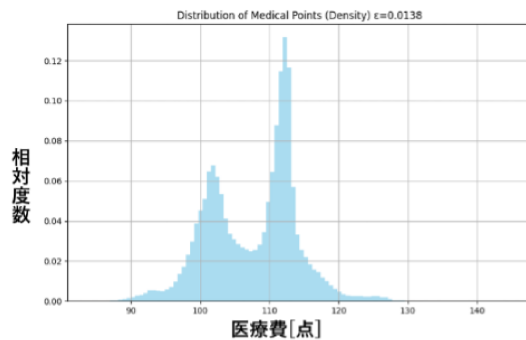


図 10 DP-XGBoost(Total $\epsilon = 0.0138$) で予測した年間医療費のヒストグラム

4. 考察

本研究では、56～65 歳の集団において、平均歩数と年間医療費の間に負の関連性が一貫して観察された。この結果は、マクロアプローチによる集計レベルの分析と、ミクロアプローチによる個票レベルの擬似データ分析の両方で支持されており、中高年層における日常的な身体活動の促進が医療費抑制に寄与する可能性を強く示唆するものである。しかしながら、この明確な関連性は、より若い年齢層においては確認されなかった。若年層の医療費は、生活習慣病よりもむしろ急性疾患や感染症など、歩数との直接的な関連が薄い要因に起因する割合が高齢層と比較して高いことも想定される。したがって、若年層においては、歩数という単一の生活習慣指標が医療費全体に与えるインパクトが相対的に小さくなっている可能性がある。これらの仮説を検証し、年齢層ごとの健康課題と医療費構造の関係をより精緻に解明するため、現在、追加の分析を進めている。

ミクロアプローチの回帰モデルを DP-XGBoost に置き換えた。その結果、3.4 に示した通り、医療費の予測値が実測と大きく乖離した。このため、統合分析の前提となる予測精度が確保できず、統合結果への影響を評価できなかった。今後は、想定以上の精度低下の原因と回避策を特定するため、今回利用した DP-XGBoost のアルゴリズム検証や、他の差分プライバシー対応回帰モデルの調査・再実験等を進めたい。

本研究で採用した統合方法は、共通属性を用いて未観測変数を回帰的に代入する「部分データ合成 (Partially Synthetic Data)」[11] に分類できると考えている。データ全体を合成する完全データ合成に関しては、データ合成手法のアルゴリズム検証や、有用性・安全性の評価手順の整備が進行しつつある。一方、部分データ合成に関しては十分に進んでいるとは言い難い。本研究が提示する統合分析のユースケースとその結果を、部分データ合成の安全性と有用性を検討する際の参考事例とすることで、議論を深めたいと考えている。

5. おわりに

本研究は、共通属性を用いたマクロアプローチとミクロアプローチによる異機関ヘルスケアデータの統合分析手法を提示し、その適用可能性と課題を示した。今後は、例えば年齢層別に適した身体活動指標や医療費抑制策をどう設定するかに加え、ミクロアプローチにおける精度低下の要因や改善策を探りつつ、本研究の事例を踏まえて部分データ合成の安全性と有用性に関する議論を深めていきたい。

参考文献

- [1] Watts, E. L., Saint-Maurice, P. F., Doherty, A., Fensom, G. K., Freeman, J. R., Gorzelitz, J. S., Jin, D., McClain, K. M., Papier, K., Patel, S. et al.: Association of accelerometer-measured physical activity level with risks of hospitalization for 25 common health conditions in UK adults, *JAMA network open*, Vol. 6, No. 2, pp. e2256186–e2256186 (2023).
- [2] Epstein, C. M. and McCoy, T. P.: Linking electronic health records with wearable technology from the All of Us Research Program, *Journal of Obstetric, Gynecologic & Neonatal Nursing*, Vol. 52, No. 2, pp. 139–149 (2023).
- [3] Leulescu, A. and Agafitei, M.: Statistical matching: a model based approach for data integration (2013).
- [4] 小暮厚之, 小林凌雅: 健康保険データに基づく医療費予測モデリング—正則化 two-part モデルによるアプローチ—, 日本保険・年金リスク学会誌= Journal of the Japanese Association of Risk, Insurance and Pensions: ジャリッブジャーナル, Vol. 10, No. 1, pp. 21–35 (2018).
- [5] Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794 (2016).
- [6] Grislain, N. and Gonzalez, J.: DP-XGBoost: Private Machine Learning at Scale (2021).
- [7] 株式会社 JMDC: JMDC Claims Database, JMDC Inc. (online), available from (<https://www.jmdc.co.jp/en/jmdc-claims-database/>) (accessed 2025-08-13).
- [8] 住友生命保険相互会社: Vitality, 住友生命保険相互会社 (online), available from (<https://vitality.sumitomolife.co.jp/>) (accessed 2025-08-13).
- [9] 厚生労働省健康・生活衛生局: 標準的な健診・保健指導プログラム (令和 6 年度版), 厚生労働省 (オンライン), 入手先 (https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000194155_00004.html) (参照 2025-08-13).
- [10] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, pp. 2623–2631 (online), DOI: 10.1145/3292500.3330701 (2019).
- [11] 高部 勲: 合成データの考え方に基づく公的統計疑似ミクロデータの作成方法の検討, 統計研究彙報= Research memoir of the statistics/総務省統計研修所 編, No. 79, pp. 111–129 (2022).