明治大学大学院　先端数理科学研究科

2023 年度

修士学位請求論文

医療データ向けの合成アルゴリズムの有用性と安全性の評価

学位請求者　先端メディアサイエンス専攻

進藤　翔太

目 次

第1章　序論

1.1　本研究の背景

医療データはより進んだ医療の提供や新たな治療法・新薬の開発, 病気の罹患予測などに活用され, 私たちがより健康で豊かな生活を送るためには欠かせない重要なデータである． 医療データは特定の個人が識別できないようにデータを加工することで，匿名加工情報として提供され，データを保有する医療機関以外でも利活用が進んでいる．例えば，ある生命保険会社は匿名加工された健診データ等を利用して, 分析した結果に基づいた保険商品を販売している [1]．被保険者は, 保険会社が提供しているアプリ上で，健診データやウェアラブルデバイスやスマートフォンを通じて取得したライフログ (歩数・心拍数等) 提供することで，生命保険会社の提供する健康増進を促すプログラムに参加することができる．その一方で，医療データは身体的な特徴・病歴などの機微な個人情報とプライバシー情報を多く含むため，第三者にデータを提供する際には個人情報保護技術を施し，安全な形で第三者に提供する安全性と，加工後も加工前と同じ質の解析ができるかという有用性のトレードオフを考慮する必要がある．

医療情報におけるプライバシーリスクとして, 2000 年のマサチューセッツ州が公開した独自に匿名化処理した医療データから州知事の病気が特定された事件 [2] が挙げられる．マサチューセッツ州は氏名のみを削除した性別, 生年月日, 郵便番号, 診療結果などの項目が含まれた医療データを研究目的で公開していた．この医療データと公開されていた投票者名簿情報から, 州知事の性別, 生年月日, 郵便番号の情報を抜き出し, マッチングさせることで条件にあう 1 人を特定し, 医療情報が推定されてしまった． この事例から，単に氏名を削除するような簡単な処理のみでは不十分であることが認識され，適切な匿名化処理を施し，安全にデータを提供することが求められるようになった．

しかし，従来の匿名化処理は大きくデータの有用性を損ねてしまう懸念がある．そこで，注目されているのが合成データである．合成データは，合成アルゴリズム [3][4][5] によって生成モデルを構築し, 生成モデルから合成された元のデータに類似する特性を持つ新たなデータのことである．この合成データを用いることで，安全に元のデータを同じような分析や精度の高い機械学習モデルを構築することができるのではないかと期待されている．

1.2　従来の合成データの課題

従来の合成手法 PrivBayes[5] で合成医療データを作成するには，次の 2 点の課題がある．

1 点目は合成データの不確実性である．合成アルゴリズムを用いて作成された合成データが，常に同じ性質を持つとは限らないことである．例えば，1 回目に合成したデータと，2 回目に同じ合成アルゴリズムで合成されたデータは異なる性質を保持する可能性がある．Theresa らは, PrivBayes な

どの合成アルゴリズムを用いて実験を行った結果，合成データがどのような特徴を保持するか予測困難であり，透明性に欠ける点を指摘している [6].

2 点目は医療データ特有の有用性があることである．医療データにおいて，どの特徴が保持されているかわからないデータを出力してしまうのは致命的である．なぜなら，こうしたデータだと特定の病気の要因などが正しく分析できず，データ利用者に間違った解釈を与えてしまう恐れがあるからだ．例えば，生命保険会社が被保険者の歩行や運動などの身体活動によって，今後の健康状態を予測し，それに応じて保険料を割り引くために医療データを使いたいという例を考える．元の医療データでは，身体活動が多いほど，特定の疾病になりにくくなるという解析がされたとする．しかし，合成されたデータでは最低限度の身体活動をしていれば，それ以上たくさん運動してもあまり疾病の予防にならないという解析結果になった場合，元データと合成データで解釈が変わってくる．生命保険会社は本来，身体活動が多くなるにつれ保険料を割り引かなければならないが，合成データを用いた場合，最低限度の身体活動をしている人すべてに保険料を一律に割り引くことになる恐れがある．その結果，生命保険会社は今後の健康状態が不安視される人を含め，必要以上に多くの人の保険料を割り引いてしまい，多額の損失を被る可能性がある．また，被保険者も生命保険会社の割引方針に応じて日常の運動方法が変わる可能性がある．仮に最低限度の運動で一律に保険料を割り引くとなっていた場合，被保険者は最低限の運動をすれば健康だと誤認し，誤った運動習慣が普及されて，健康リスクを上げてしまう恐れがある．つまり，この例では合成データが正しく病気と因子の関係がとらえられていないことにより，保険会社では経済的なリスク，被保険者は健康リスクを上げてしまうことになり，双方が不利益を被る結果となる．

PrivBayes では，目的変数を指定してデータを合成することができず，病気と因子の関係を具体的に捉えることができない．そのため，有用性の高い医療データを合成するのが困難である．

1.3 研究目的

そこで，PrivBayes の課題に対し，本研究では病気とその因子の特徴を捉え，データ活用者が正しく病気の原因を分析可能な医療データ向け合成アルゴリズムを提案することを目的とする．

1.4 研究の新規性

本研究ではこの問題に対して，ロジスティック回帰モデルに着目し，医療データ向けの合成アルゴリズムを提案する．ロジスティック回帰は医療統計で広く用いられる代表的な統計手法である．複数の要因から病気の発生確率を予測し，オッズ比を求めることで交絡因子を排除して特定の因子がどの程度，病気に影響を及ぼしているかを分析できる．PrivBayes では目的変数 (病気) を指定してデータを合成していないため，病気と因子の関係を上手に捉えることができておらず，医療データとして有用性が低い合成データを出力してしまっている．

提案手法は，説明変数（因子）は PrivBayes で合成し，目的変数（病気）はロジスティック回帰で推定した確率 p に従って疾病の罹患を合成する．それによって，因子と病気の関係を正確に捉えて

データを合成し，データ活用者がより正しく病気の原因を調査することができる医療合成データを作成する．

提案手法は，PrivBayes のみならず，CTGAN[3]，Gaussian copula[4] などの他の合成アルゴリズムにも適用可能である．そこで，本研究では PrivBayes のみならず，CTGAN，Gaussian copula に提案手法を適用した場合についても有用性と安全性の評価を行う．既存のアルゴリズムと提案手法で合成されたデータでそれぞれ，ロジスティック回帰の有用性評価 (病気の因子分析)，機械学習モデルの有用性評価を行い，データ活用者にとって有用なデータを合成できているかを実験的に評価する．また，合成データから特定のセンシティブな値を推定する攻撃，GCAP[13] を用いて合成されたデータの安全性を実験的に評価する．

1.5 本研究の貢献

本研究の貢献は以下の 3 つである．

- ロジスティック回帰モデルに基づく，病気と因子の関係を捉えた医療データ向けの新しい合成アルゴリズムを提案したこと．
- 合成データにおいてロジスティック回帰 (病気の因子分析) の精度を評価する独自の有用性評価指標を提案したこと．
- 提案手法の安全性の評価を行い，提案方式が既存手法よりも高い有用性を持つことを示したこと．

1.6 本稿の構成

本稿の構成は以下のようになっている．

- 1 章：本研究の背景と目的について述べた．
- 2 章：本稿の基本定義と従来研究について述べる．
- 3 章：医療データの活用事例について述べる．
- 4 章：医療データ向けの合成アルゴリズムの提案と有用性評価指標の提案を行う．
- 5 章：有用性と安全性の評価を行う．
- 6 章：本稿のまとめを行う．

第2章　基本定義と従来研究

2.1　基本定義

2.1.1　ロジスティック回帰分析

ロジスティック回帰分析は医療統計で多く用いられる統計手法である．糖尿病に罹患する確率を p とすると，

$$p = \frac{1}{1 + e^{-(a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_i x_i)}} \tag{2.1}$$

で表される．ここで，変数 x_i には, 年齢，BMI，喫煙など様々な因子が含まれる．a_i は各変数の係数, a_0 は定数である．ロジスティック回帰分析は交絡因子の影響を調整して，各因子がどの程度，糖尿病に影響を及ぼしているかを調べることができる．各因子の影響度を測る指標として知られているオッズ比 OR(Odds Ratio) は

$$OR = e^{a_i} \tag{2.2}$$

で定義される．

ここで，x_i 以外の説明変数が同じ値で, x_i のみを 1 増加させた時の罹患する確率を q とすると，それぞれのオッズは

$$OR = \frac{p}{1-p} = e^{a_1 x_1 + a_2 x_2 + \cdots + a_i x_i + a_0} \tag{2.3}$$

$$OR|_{x_i \leftarrow x_i + 1} = \frac{q}{1-q} = e^{a_1 x_1 + a_2 x_2 + \cdots + a_i (x_i + 1) + a_0} \tag{2.4}$$

と表されるので，この比をとると，

$$OR = \frac{q}{1-q} \Big/ \frac{p}{1-p} = e^{a_i} \tag{2.5}$$

になり，特定の変数が 1 増加した時にオッズが何倍になるかを表している．

また，ある組織の喫煙者と糖尿病患者数が表 2.1 の集計表において，非喫煙者を基準とした喫煙者の OR は,

$$OR = \frac{a/b}{c/d} \tag{2.6}$$

で定義され，相対リスク RR(Relative Risk) は

$$RR = \frac{a/a+b}{c/c+d} \tag{2.7}$$

で定義される．従って，$a \ll b$，$c \ll d$ の場合，$OR \approx RR$ となり，オッズ比は相対リスクを近似するとみなせる．

表 2.1: 喫煙と糖尿病の集計表

	糖尿病である	糖尿病でない	計
喫煙者	a	b	$a+b$
非喫煙者	c	d	$c+d$
計	$a+c$	$b+d$	N

2.1.2 属性推定攻撃

医療データには，個人を特定しうる情報や公開されたくないプライバシーに関わる情報が多く含まれている．マサチューセッツ州の例 [2] では，氏名のみを削除した医療情報を公開したことにより，性別や生年月日，郵便番号から，知事の病気や請求額などの情報が特定されてしまった．このように組み合わせることで個人を特定しうる性別や生年月日，郵便番号のような属性のことを準識別子と呼び，病気や請求額のようにプライバシーに関わり公開されることが望ましくない属性のことをセンシティブ属性と呼ぶ．図 2.1 の例で説明すると，攻撃者は医療情報を用いて，特定の個人の診断結果と請求額 (センシティブ属性) を抜き出したいと仮定する．しかし，この医療情報には氏名が無く，氏名を参照して個人を特定することはできない．そこで，攻撃者は事前に所持していた性別，郵便番号，生年月日 (準識別子) を用いて，個人を識別し，診断結果と請求額 (センシティブ属性) を抜き出す．このように，準識別子を用いてセンシティブ属性を推定する攻撃を属性推定攻撃と呼ぶ．

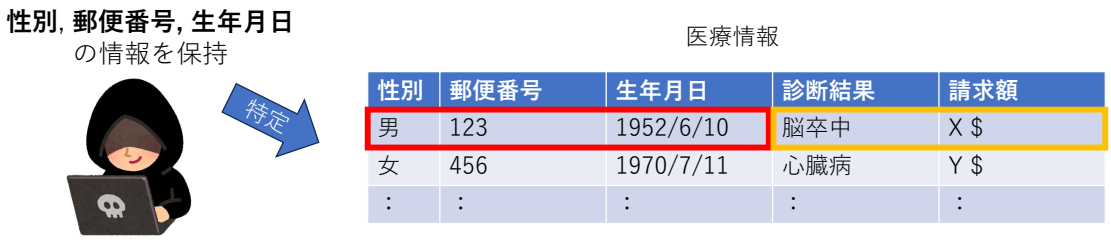


図 2.1: 属性推定攻撃の例

2.1.3 差分プライバシー

差分プライバシー [7] は，Dwork によって考案されたプライバシー保護の統一的な安全性指標のことである．

あるメカニズム $M : D \to S$ において，1 つの要素のみが異なる任意のデータセット D，D' と，任意の出力の集合 $S \subset range(M)$ について

$$Pr[M(D) \in S] \leq e^{\epsilon} \cdot Pr[M(D') \in S]$$

を満たすとき，メカニズム M は ϵ-差分プライバシーを満たす．

例えば，データセット D' は，D の 1 つのレコードを (x_i, y_i) とすると, そのレコードのみが (x'_i, y'_i) に置き換わったデータセットのことを指す．ϵ が小さいほど，メカニズム M に D と D' を入力して，返ってきた出力結果 $M(D)$ と $M(D')$ の差が小さくなるため，個人 (1 つのレコード) のプライバシーが保護され，安全性が保証される．つまり，差分プライバシーの定義では，小さい ϵ を満たすメカニズム A ほど安全性が高くなる．

2.2 従来研究

2.2.1 合成アルゴリズム

PrivBayes

PrivBayes[5] はベイジアンネットワークをベースとした合成モデルである．PrivBayes はベイジアンネットワークの構造を決定するフェーズと，条件付き確率を求めるフェーズの二段階からなる．ベイジアンネットワークの構造は相互情報量を用いて決定する．Algorithm1 により，相互情報量の最も大きい親と子の組み合わせで，グラフ構造を作成していく．子のノードは次の親の候補に回るため，最終的なグラフ構造は有向非巡回グラフ (DAG) となる．相互情報量はノード X と親の候補を Ω とすると，次のように定義される．

$$I(X, \Omega) := \sum_{\substack{x \in dom(X) \\ x \in dom(\Omega)}} Pr[(X, \Omega) = (x, \omega)] \log \frac{Pr[(X, \Omega) = (x, \omega)]}{Pr[X = x] Pr[\Omega = \omega]}$$

これは，X と Ω が独立している場合 $I(X, \Omega) = 0$ になり，X と Ω の相関が強く，互いに影響を及ぼし合っているほど，値は大きくなる．つまり，相互情報量は 2 つの属性間の依存性を表している．

Algorithm1 でベイジアンネットワークの構造が決定したら，条件付き確率を計算する．例えば，親ノードが人種 race で，子ノードが糖尿病 dia という構造になっている場合の条件付き確率 $Pr[dia|race]$ を考える．この条件付き確率は，値の頻度から算出される．例えば，$Pr[dia = 1|race = Black]$ を求めるには，表 2.2 のように，元データから race と dia の集計をし，

$$Pr[dia = 1|race = Black] = \frac{Pr[dia = 1, race = Black]}{Pr[race = Black]} = \frac{200/1800}{600/1800} = \frac{1}{3}$$

と求める．

また，PrivBayes は ϵ-差分プライバシーを保証して，合成モデルを作成し，合成データを出力することができる．Algorithm2 のように，同時確率分布に対して，ϵ によって決まるラプラスノイズを加えてから，条件付き確率を求めることで，合成アルゴリズムの ϵ-差分プライバシーを保証する．差分プライバシーの定義では ϵ の値が小さいほど安全性が保証されるため，ϵ が小さくなるほど加えるラプラスノイズが大きくなる．

表 2.2: 人種と糖尿病の集計表

	糖尿病 (dia=1)	糖尿病でない (dia=0)	計
Black	200	400	600
White	300	900	1200

Algorithm 1 Greedy Bayes [5]

Input: D：データセット，k: 親ノードの最大数

Output: G: グラフ構造

1: $V = \phi$
2: ルートノードとなる属性 X_1 を 属性の集合 A からランダムに決め，X_1 を V に追加する．
3: **for** $i = 2, \ldots, d$ **do**
4: $\quad \Omega = \phi$
5: $\quad V$ の中の k 個のノードの組み合わせ全てを Ω に入れる．
6: $\quad$ **for** $X_i \in A \backslash V$ **do**
7: $\quad\quad$ 任意の $\Omega_i \in \Omega$ と X の相互情報量 $I(X_i, \Omega_i)$ を計算する．
8: $\quad$ **end for**
9: $\quad$ 最も相互情報量が大きい X_i と Ω_i を取り出し，子ノードを X_i，親ノードを Ω_i とし，G に追加する．X_i を V に追加する．
10: **end for**

Gaussian copula

Gaussian copula[4] は統計をベースにした合成モデルである．コピュラは変数間の依存関係を示す関数のことであり，同時確率分布と周辺分布をつなぎ合わせる．データセットの列を 0, ..., n をとし，それぞれの累積分布の逆関数を $F_0^{-1}, ..., F_n^{-1}$ とする．また，同時分布を $H(x_0, ..., x_n)$ とすると，以下の関係が成り立つ．

$$H(x_1, ..., x_n) = C(F_0^{-1}(x_0), ..., F_n^{-1}(x_n)) \tag{2.8}$$

C がコピュラであり，F_0, ..., F_n が連続であるとき，コピュラ C は一意に定まる．Gaussian copula は周辺分布をガウス分布であると仮定し，合成データを出力する．

Algorithm 2 Noisy Conditionals [5]

Input: D：データセット，G：グラフ構造，k: 親ノードの最大数

Output: P^*: ノイズを加えて推定した条件付き確率分布

$P^* = \phi$
for $i = k+1, \ldots, d$ **do**
　同時確率分布 $Pr[X_i, \Omega_i]$ を計算する．
　$Pr^*[X_i, \Omega_i] \leftarrow Pr[X_i, \Omega_i] + Lap(\frac{2(d-k)}{n\epsilon_2})$
　$Pr^*[X_i, \Omega_i]$ が負の成分を 0 にし，全体が 1 になるように正規化する．
　ノイズを含んだ同時確率分布 $Pr^*[X_i, \Omega_i]$ から条件付き確率分布 $Pr^*[X_i|\Omega_i]$ を計算し，P^* に加える．
　for $j = 1, \ldots, k$ **do**
　　ノイズを含んだ k+1 次元の同時分布 $Pr^*[X_{k+1}, \Omega_{k+1}]$ から条件付き分布 $Pr^*[X_j|\Omega_j]$ を計算し，P^* に追加する．
　end for
end for

CTGAN

CTGAN[3] は深層学習をベースとした合成モデルである．モデルは，データを生成する Generator とそれが合成のデータであるかを判別する Critic の 2 つのニューラルネットワークで構成される．Generator は，Critic を騙すように合成データを出力するモデルを学習していき，Critic はそれを正しく合成と見抜くように識別器を学習していく．つまり，Critic と競い合わせることで，本物に近いデータを出力する合成モデル Generater を作る仕組みである．

2.2.2 属性推定の安全性評価指標

CAP

CAP(Correct Attribution Probability)[12] はレコード一致をベースとする属性推定攻撃である．CAP では攻撃者が元データに含まれる個人の主要な属性であるキー属性と，合成データを使って推定したいターゲット属性を推定する．n レコード (n 人)，k 属性 (列) の元データセットで，$j \in \{1, ..., n\}$ とし，$K_{o,j}$ を j 番目の元データセットのキー属性，$T_{o,j}$ をターゲット属性とする．同様に，n レコードの合成データセットで，$i \in \{1, ..., n\}$ とすると，レコード i のキー属性，ターゲット属性をそれぞれ $K_{s,i}$，$T_{s,i}$ で表す．CAP は $K_{o,j}$ が与えられた時の $T_{o,j}$ の条件付き確率，すなわち，

$$CAP_{s,j} = P_s(T_{o,j} \mid K_{o,j}) = \sum_{i=1}^{n} \frac{|\{T_{s,i} = T_{o,j} \cap K_{s,i} = K_{o,j}\}|}{\{K_{s,j}, K_{o,j} \in \{1, ..., k\} | K_{s,i} = K_{o,j}\}} \tag{2.9}$$

と定める．個人 j のキー属性と一致する合成データのレコードを取り出し，そのうちターゲット属性が $T_{o,j}$，すなわち，元データと同一のレコードの割合で確率推定を行う．

GCAP

しかし，CAP ではターゲットのキー属性の組み合わせが合成データに存在しなかった場合には確率を計算することができない．仮にターゲットのキー属性と全て一致できなかった場合にも，確率を計算できるように一般化したものが GCAP[13] である．GCAP はハミング距離を使って確率を計算する．ハミング距離は，個人 j のキー属性を $K_{o,j} = (a_1, ..., a_\ell)$，合成データの i 番目のレコードのキー属性を $K_{s,i} = (b_1, ..., b_\ell)$ とすると，

$$\Delta(K_{o,j}, K_{s,i}) = |\{k \in \{1, ..., \ell\} | a_k \neq b_k\}| \tag{2.10}$$

で定義される．例えば，全てのキーが一致するときは $\Delta(K_{o,j}, K_{s,i}) = 0$，1 つだけ一致しないときは $\Delta(K_{o,j}, K_{s,i}) = 1$ となる．GCAP ではまず，以下のようにまずハミング距離の最小値 ρ を求める．

$$\rho = \min\{r \mid i \in \{1, ..., n\}, \Delta(K_{o,j}, K_{s,i}) = r\} \tag{2.11}$$

そして，GCAP はハミング距離が最小値 ρ となる合成データのレコードを参照することで，

$$GCAP_{s,j} = \sum_{i=1}^{n} \frac{|\{T_{s,i} = T_{o,j} \cap \Delta(K_{o,j}, K_{s,i}) = \rho\}|}{\{\Delta(K_{o,j}, K_{s,i}) = \rho\}} \tag{2.12}$$

とターゲット属性の推定確率を求める

第3章　医療情報研究

3.1　コホート研究

医療情報を用いた研究は数多く行われてきた．1つ代表的なものとして挙げられるのはコホート研究である．コホート研究とは，特定の疾病の発生要因などを調べるために，疾病に関わるであろう要因を持つ集団とそうした要因を持っていない健康な集団に分け，一定期間の追跡調査を行う観察的研究の手法の1つである．

Pedroらは歩数と死亡リスクによるコホート研究を行った[8]．米国・国立衛生研究所の研究グループは，歩行数と死亡の要因を調べるために，米国健康栄養調査 (National Health and Nutrition Survey) に参加した4,840人を対象に10年以上追跡調査を行った．参加した4,840人の歩数を追跡前に7日間ほど歩数計で計測した後，歩数が4,000歩のグループと8,000歩のグループに分け，それぞれで10年間で死亡した割合や交絡因子を調整して死亡リスクを計算した．調査の結果，8,000歩のグループは4,000歩のグループに比べ，死亡リスクが半減していることが明らかとなり，統計的にも有意な結果であった．

しかし，このコホート研究を行うためには，データの収集が困難な点がある．被験者を大規模に集めデータを収集しなければならかったり，歩数などを長期的に計測するコストが高い．実際に，この研究では歩数を7日間しか計測できていない．

3.2　匿名加工医療データの応用例と優位性

コホート研究の例から，医療データを用いた研究において，大きな課題となるのは大規模にデータを収集することにあると言える．匿名加工医療データはそうしたデータ収集の課題を解決する手段となりうる．

匿名加工医療データの応用例として大阪大学のチームが行った糖尿病・脂質異常症・高血圧の3つの生活習慣病の3年以内の発症を予測するAIを作成した研究が挙げられる[9][10]．この研究においても，データ収集が課題であった．生活習慣病の発症リスクを正確に予測するAIモデルを構築するためには，健診結果などの個人情報を多く含む医療データが必要であり，なかなか十分なデータ数が集まらないという課題があった．著者らがこのデータ収集の困難さを解決した手段が，匿名加工医療データである．この研究は，大阪府の各市町村における国民健康保険の被保険者の特定検診データを匿名加工を施した，匿名加工医療データを用いて行われた．このデータには54万人分の約6年分の健診データが含まれており，大阪府健保連合会などの協力によって大阪大学のチームに提供された．

研究の結果，大量のデータを使って予測AIを構築する際の優位性を実証し，高い精度で生活習慣病を予測するAIモデルを構築した．開発された予測モデルは大阪府が運営する健康支援アプリ，ア

スマイルに組み込まれており，人々の健康促進や生活習慣改善の手助けとなっている．

3.3　医療データの利用目的と課題

コホート研究の事例と，匿名加工医療データの応用事例から医療データの利用目的は主に 2 つであると考える．1 つ目は，コホート研究などのように統計手法などを用いて，特定の疾病の原因を分析こと．2 つ目は，機械学習などを用いて病気の予測モデルを構築し，特定の疾病の罹患予測を行うことである．

一方で，医療データには身体的な特徴・病歴などの機微な個人情報とプライバシー情報を多く含むため，プライバシーの懸念がつきまとう．本研究では，医療データを合成データとすることでこの課題について解決することを目指す．医療データに向けの合成アルゴリズムを提案し，プライバシーの懸念を排除して，特定の疾病に対する原因の解析と特定の疾病の罹患予測モデルを構築可能な医療データの合成を試みる．

第4章　医療データ向け合成手法と有用性評価の提案

4.1　従来の有用性指標

データの有用性を測る際にMAE（Mean Absolute Error）が広く用いられている．Theresaらは合成データの有用性と安全性の評価を行ったが，有用性をMAEで評価した[6].

例えば，合成データのロジスティック回帰のオッズ比について，元データを$orig$，合成データをsyn，オッズ比をOR，オッズ比の各要素をiとすると，MAEは，

$$U_1 = \sum_{i=1}^{n} \frac{1}{n} | \, OR_{orig,i} - OR_{syn,i} | \tag{4.1}$$

と求まる．これは，合成データにおけるオッズ比の誤差の平均を計算している．すなわち，U_1が0に近づくほど有用性が高い.

4.2　MAEの問題点

しかし，MAEのみを用いて，ロジスティック回帰のオッズ比における有用性を十分に測ることはできない．例えば，元データを使って，糖尿病に対するMETs(運動の強度)のオッズ比を求めた表4.1を考えよう．METsは運動強度を表していて，METsがかなり低い時を基準とした各項目のオッズ比を求めている．影響度の順位は糖尿病になりやすい項目の順位，つまりオッズ比が高い順番を示している．一方，表4.2と表4.3はそれぞれ違う合成データから求めたオッズ比の例である．2つともMAEで評価すると，同じ有用性になる．しかし，本当に有用性が同じと言えるだろうか？

例えば，それぞれの影響度の順序を見てみると，表4.2はMETsが高くなるほど影響度の順位が上がっていくため，高い強度の運動をしている人ほど糖尿病になりやすいと解釈される．一方，表4.3はMETsが高くなるほど影響度の順位が下がっていくので，高い強度の運動をしているほど糖尿病になりやすいと解釈される．これは，元データの順位と同じである．つまり，結果の整合性という点では表4.3の方が有用性が高いと言える．しかし，MAEでは単純なオッズ比の誤差しか見ていないため，結果に整合性のある表4.3の結果と，元データと全く逆の解釈になってしまう表4.2で同じ有用性という評価になってしまう．これでは，不十分であり，元データとの結果の整合性を評価できるような有用性評価を加える必要がある.

4.3　順位相関を用いたオッズ比の有用性指標の提案

MAEだけでは結果の整合性を評価できないため，整合性を評価できる有用性指標を提案する.

表 4.1: 元データの OR(METs の項目)

METs	オッズ比	影響度の順位
高い	0.7	4
中程度	0.8	3
低い	0.9	2
かなり低い	1	1

表 4.2: 合成データ 1 の OR

METs	オッズ比	影響度の順位
高い	1.3	1
中程度	1.2	2
低い	1.1	3
かなり低い	1	4

表 4.3: 合成データ 2 の OR

METs	オッズ比	影響度の順位
高い	0.3	4
中程度	0.4	3
低い	0.5	2
かなり低い	1	1

n 行 K 列のデータを合成する．目的変数以外の各 K 個の属性 i について，オッズ比が大きい順に並べて，影響度のランク $rank^{(i)}$ を付与し，元データと合成データのランク，$rank_{orig}$ と $rank_{syn}$ についてのスピアマンの順位相関係数

$$U_2 = \sum_{k=1}^{K} \frac{1}{K}(1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (rank_{orig,k}^{(i)} - rank_{syn,k}^{(i)})^2) \tag{4.2}$$

を合成データの有用性 U_2 と定義する．例えば，元データと結果の解釈が同じになる表 4.3 のような場合には，$U_2 = 1$ となり，元データと結果の解釈が逆になってしまう表 4.2 のような場合には，$U_2 = -1$ となる．すなわち，$U_2 = 1$ に近い合成データほど有用性が高く，$U_2 = -1$ に近い合成データほど有用性が低くなる．

4.4 PrivBayes の課題

PrivBayes[5] の課題として，目的変数を指定して合成できないため，出力される合成データの特性がランダムに変わってしまうことが挙げられる．PrivBayes はベイジアンネットワークをベースとした合成アルゴリズムである．PrivBayes では，ベイジアンネットワークを組む際に，最初のノードをランダムで決める．

実際に NHANES データセットで，PrivBayes を用いて合成モデルを構築した際の例を挙げる．最初のノードに目的変数 *dia* が選ばれたときと，それ以外のノードが選ばれた時の *dia* に関するベイジアンネットワークをそれぞれ図 4.1，図 4.2 に示す．図 4.1 は *dia* から *age*, *mar* を合成するモデルを構築しているのに対し，図 4.2 では *age*，*bmi* から *dia* を合成するモデルを構築している．それぞれに関わる説明変数において，*age* は共通しているが，図 4.1 では *mar*，図 4.2 では *bmi* と全く別の変数が関わっており，それぞれの合成モデルで捉えている特徴が違う．そのため，それぞれのモデルで出力されるデータの性質が変わってしまう．例えば，最初のノードを変えて，PrivBayes でそれぞれ

10 回合成モデルを生成し，元データと同じサンプルサイズでサンプリングした時のロジスティック回帰のオッズ比の誤差 MAE を図 4.3 に示す．提案方式による順位相関を用いた結果の整合性の評価を図 4.4 に示す．オッズ比の MAE は最初のノードの選び方によってばらつきに差があり，不確実性が高い．従来方式のオッズ比の順位誤差は目的変数である *dia* を最初のノードにしたとき，極端に提案する有用性が小さくなっている．つまり，他のノードに比べ，結果の整合性を担保せず，医療データとして糖尿病の分析モデルを作ることが困難である．

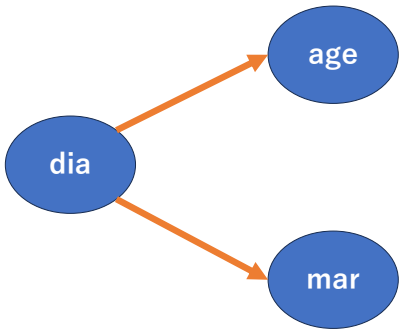


図 4.1: dia を最初のノードとした時のネットワーク

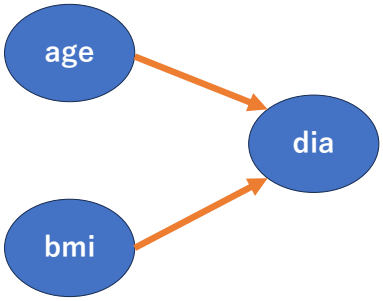


図 4.2: dia 以外を最初のノードとした時のネットワーク

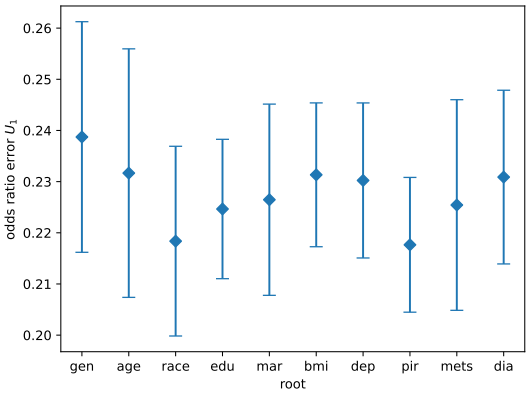


図 4.3: 最初に選ばれたノードごとのオッズ比の誤差 (MAE)

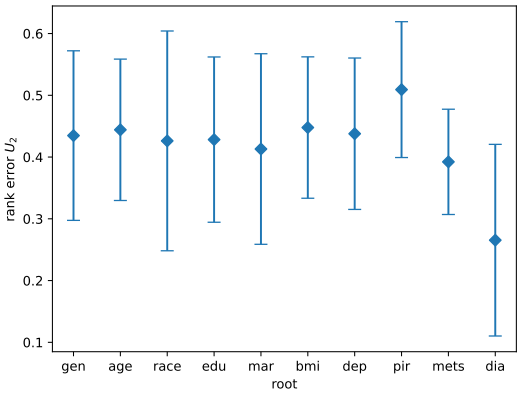


図 4.4: 最初に選ばれたノードごとの提案する有用性 (オッズ比の順位誤差)

4.5 医療データ向けの合成手法の提案

既存の合成アルゴリズムの課題を解決するため，本研究では目的変数を指定し，説明変数と目的変数を分けて合成する手法を提案する．提案手法のアルゴリズムを Algorithm3 に示す．

合成手順について説明する．1. 元データから説明変数 $X_1, \ldots, X_{d-1}$ のみを取り出し，説明変数を合成する合成モデル M を作成する．2. 作成した合成モデル M で説明変数 $\widehat{X_1}, \ldots, \widehat{X_{d-1}}$ を合成する．3. 元データからロジスティック回帰モデル f を作成し，回帰モデルに合成した説明変数の i 番目のレコード $\widehat{x_{i,j}} \sim M$ を f に代入して確率 $p = Pr[\widehat{x_{i,d}} = 1 | \widehat{x_{i,1}}, \ldots, \widehat{x_{i,d-1}}]$ を計算する．4. 確率 p にしたがって目的変数を生成する．

Algorithm 3 ロジスティック回帰を用いた合成手法

Input: D_{orig}：データセット，X_d：目的変数の指定，s：出力データサイズ

Output: D_{syn}：s 行の合成データセット

1: $D_{syn} = \phi$

2: 説明変数 $X_1, \ldots, X_{d-1} \in D_{orig} \backslash X_d$ で合成モデル M を作成する．

3: 合成モデル M で s 行の説明変数 $\widehat{X_1}, ..., \widehat{X_{d-1}}$ を合成する．

4: 説明変数 $X_1, \ldots, X_{d-1} \in D_{orig} \backslash X_d$，目的変数 X_d として，ロジスティック回帰モデル f を作成する．

5: **for** $i = 1, \ldots, s$ **do**

6: 　合成された説明変数の i 行目のレコード $\widehat{x_{i,1}}, \ldots, \widehat{x_{i,d-1}}$ をロジスティック回帰モデル f に代入して，目的変数の予測確率 $p = Pr[\widehat{x_{i,d}} = 1 | \widehat{x_{i,1}}, \ldots, \widehat{x_{i,d-1}}] = f(\widehat{x_{i,1}}, \ldots, \widehat{x_{i,d-1}})$ を計算する．

7: 　確率 p に従い $\widehat{x_{i,d}}$ の値を決定する．$\widehat{x_{i,d}} \begin{cases} 1 & with\ \ prob\ \ p \\ 0 & with\ \ prob\ \ 1-p \end{cases}$

8: 　D_{syn} に $\widehat{x_{i,1}}, \ldots, \widehat{x_{i,d}}$ を追加する．

9: **end for**

例えば，表 4.4 の合成モデル M で説明変数 X_1, X_2, X_3 のみを合成した例を考える．この合成データは目的変数 X_d のみが作られていない．X_d はロジスティック回帰モデル f に説明変数 X_1, X_2, X_3 を代入して，確率 p を求めることで作成する．1 行目のレコードで，説明変数を代入して $p = f(X_1 = Male, X_2 = Black, X_3 = 64) = 0.7$ となったとする．この時，1 行目のレコードは確率 $p = 0.7$ で $X_d = 1$，確率 $1 - p = 0.3$ で $X_d = 0$ と合成される．同様に，2 行目のレコードも確率 p を求め，$p = f(X_1 = Female, X_2 = White, X_3 = 50) = 0.4$ だったとする．2 行目のレコードは確率 $p = 0.4$ で $X_d = 1$，確率 $1 - p = 0.6$ で $X_d = 0$ と合成される．

表 4.4: 合成された説明変数の例

	gen(X_1)	race(X_2)	age(X_3)	dia(X_d)
1	Male	Black	64	?
2	Female	White	50	?

第5章　評価実験

5.1　データセット

5.1.1　NHANES データセット

NHANES(National Health and Nutrition Survey) は米国健康栄養調査である．米国疾病対策予防センター (CDC) は，主要な病気の疾病率とその危険因子を特定することを目的として米国で毎年 5,000 人ほどの人数をランダムにサンプリングし，被験者の持つ疾病や問診，診断結果のデータを収集し調査を行なっている．本研究では，NHANES2015-2016 のデータから，gen(性別)，age(年齢)，race(人種)，edu(学歴)，mar(婚姻状態)，bmi，dep(鬱状態)，pir(貧困状態)，mets(運動強度)，dia(糖尿病) の 10 属性取り出し，欠損値を持つレコードを排除した 4,190 人の糖尿病データセットを作成し，実験を行う．データセットの統計量を表 5.1，表 5.2 に示す．

表 5.1: NHANES 糖尿病データセットの基本統計量（離散値）

	gen	race	edu	mar	dep	pir	mets	dia
要素数	2	5	5	6	2	2	4	2
最頻値	Female	White	College	Married	0	0	Q1	0
頻度（最頻値）	2117	1398	1214	2171	3313	3306	1168	3317

表 5.2: NHANES 糖尿病データセットの基本統計量（連続値）

	age	bmi
平均値	50.455847	29.184010
標準偏差	17.887312	6.850947
最小値	20.000000	14.500000
25%	35.000000	24.400000
50%	50.000000	28.100000
75%	65.000000	32.700000
最大値	80.000000	67.300000

5.1.2　DeSC データセット

本研究では，DeSC ヘルスケアから法律に従って，適切に加工されたヘルスケアデータの匿名加工情報を利用する．本データの統計量を表 5.3 に示す．また，このデータセットを全て使用するのでは

なく，前処理を施して利用した．

前処理の手順について説明する．最初に 2017 年度の健康診断データから，表 5.5 の種別が病気以外の 30 項目を取り出し，欠損値のあるレコードを削除する．その後，健康診断データに含まれる個人のうち，レセプトデータを参照して 2014–2017 年の間に糖尿病の診断記録がある個人を削除する．2018–2020 年の三年間でレセプトデータに糖尿病の診断記録がある個人を糖尿病であるとし，それ以外の個人は糖尿病でないとラベルを付与した．前処理を施したデータセットの基本情報を表 5.4 に示す．

また，この 44,407 人全てのデータを使わず，ランダムにサンプリングした 5,000 人を訓練データ (合成アルゴリズムに学習させるデータ) とし，それ以外からランダムにサンプリングした 5,000 人をテストデータ (機械学習モデルの有用性を測るためのデータ) として利用した．

表 5.3: DeSC データセットの統計量

	レセプトデータ	健康診断データ
属性数	11	105
期間	2014–2021	2014–2021
レコード数	2,516,102,835	7,028,931
対象者	9,597,522	2,345,128

表 5.4: 前処理後のデータセットの基本情報

基本情報	前処理後
対象者	44,407
糖尿病罹患数 (割合)	8,791(0.20)
性別 (男性割合)	男性 (0.56)
年齢	20-95

5.2 有用性評価実験

5.2.1 実験方法

提案手法は，PrivBayes 以外の合成アルゴリズムにも適用可能である．そのため，PrivBayes のみならず，他の合成アルゴリズムにも適用し，評価を行う．実験に使用する合成アルゴリズムを表 5.6 に示す．PrivBayes($\epsilon = \infty$)，CTGAN，Gaussian copula の 3 つの従来手法と，PrivBayes($\epsilon = \infty$)+提案手法，CTGAN+提案手法，Gaussian copula+提案手法，差分プライバシーを保証した PrivBayes($\epsilon = 1$, $\epsilon = 0.5$, $\epsilon = 0.1$) の全 9 個の合成モデルでデータを合成し，有用性を評価する．NHANES データセットに前処理を施し作成した糖尿病データセットと，DeSC データセットに前処理を施し作成した糖尿病データセットを用いて実験を行う．本研究では次の 2 つの評価を行う．

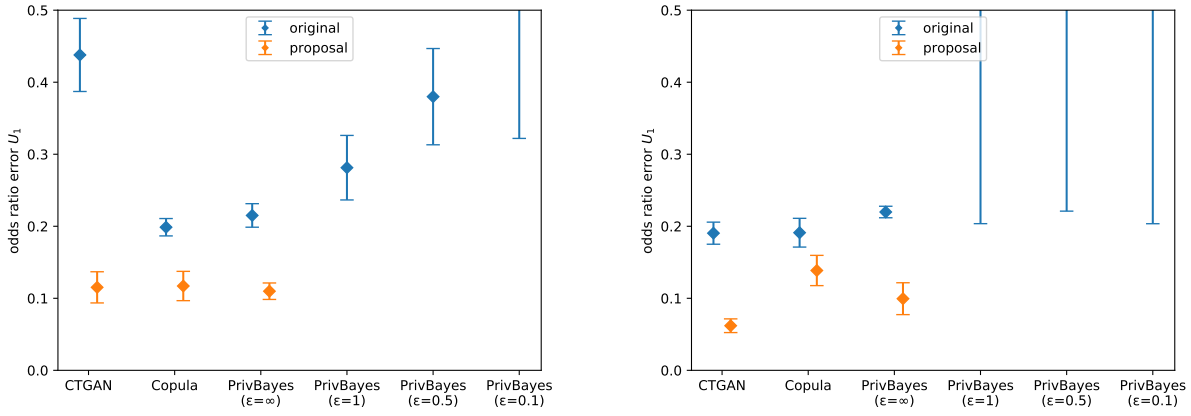
(実験 1) ロジスティック回帰の有用性　これは，3 章で触れた特定の疾病の原因を合成データで正しく分析できるかという観点での評価にあたる．元データと同じデータサイズで 10 回データを合成し，オッズ比の誤差，順位相関を用いて結果の整合性をそれぞれ評価する．

(実験 2) 予測モデルの有用性　これは，3 章で触れた合成データを用いて病気の予測モデルを構築し，特定の疾病の罹患予測を行えるかという観点での評価にあたる．予測モデルは Random Forest(RF) で予測モデルを構築し，テストデータで予測した際の F 値で評価を行う．テストデータは，DeSC データセットでは訓練データの 5,000 人以外のデータからランダムに取り出した 5,000 人のデータセットを用いる．また，NHANES データセットでは 4190 人のデータセットを半分に分け，訓練データ 2095 人，テストデータ 2095 人のデータセットで評価を行う．訓練データと同じサイズで 10 回データを合成し，予測モデルの有用性を評価する．

5.2.2　実験 1: ロジスティック回帰の有用性

図 5.1 にオッズ比の MAE を示す．CTGAN，Gaussian copula，PrivBayes の全ての合成アルゴリズムに対し，提案手法で目的変数を合成した方式は誤差が小さくなった．NHANES データセットでは CTGAN で平均 0.32，Gaussian copula で平均 0.08，PrivBayes では平均 0.10 ほど誤差が小さくなっていた．DeSC データセットにおいても，CTGAN で平均 0.13，Gaussian copula で平均 0.05，PrivBayes では平均 0.12 ほど誤差が小さい．

PrivBayes の $\epsilon = 1$，$\epsilon = 0.5$，$\epsilon = 0.1$ で差分プライバシーを満たす合成モデルでは，NHANES データセットで $\epsilon = 1$ で MAE が平均 0.28，$\epsilon = 0.5$ で平均 0.38，$\epsilon = 0.1$ で平均 4.1×10^9 と ϵ が小さくなるにつれ，有用性が下がった．同様に DeSC データセットにおいても，$\epsilon = 1$ で MAE が平均 5.5×10^8，$\epsilon = 0.5$ で平均 4.4×10^9，$\epsilon = 0.1$ で平均 1.4×10^{10} と ϵ が小さくなるにつれ，有用性が下がった．



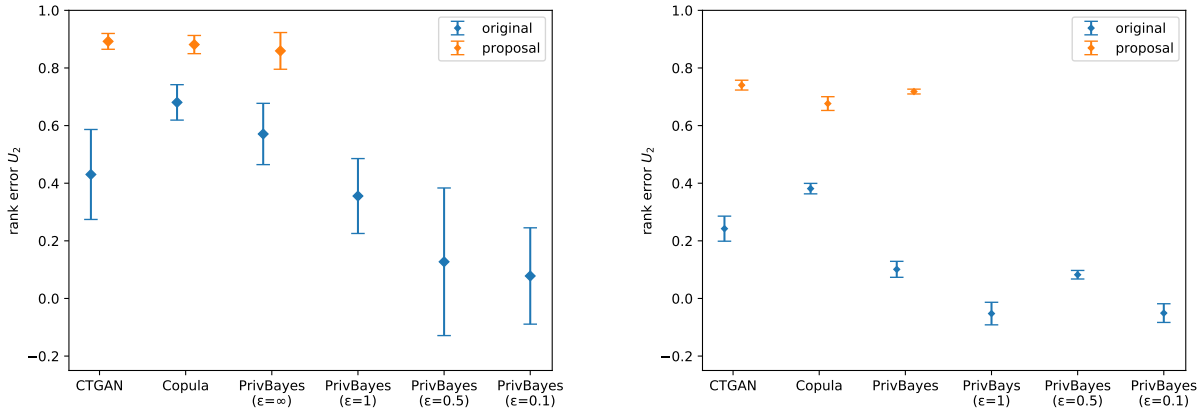
(a) NHANES データセット　　**(b)** DeSC データセット

図 5.1: 従来の有用性 (オッズ比の MAE)

図 5.2 に提案有用性であるオッズ比の順位相関を示す．オッズ比の誤差と同様にオッズ比の順位相関，すなわち，病気に対する因子の影響度の順序を保持しているかどうかについての有用性も，提案

合成方式の方が高かった．NHANES データセットでは CTGAN で平均 0.46，Gaussian copula で平均 0.21，PrivBayes では平均 0.29 ほど順位相関が高くなっていた．DeSC データセットにおいても，CTGAN で平均 0.50，Gaussian copula で平均 0.30，PrivBayes では平均 0.62 ほど順位相関が高くなっていた．

PrivBayes の $\epsilon = 1$，$\epsilon = 0.5$，$\epsilon = 0.1$ で差分プライバシーを満たす合成モデルでは，NHANES データセットで $\epsilon = 1$ で順位相関が平均 0.36，$\epsilon = 0.5$ で平均 0.12，$\epsilon = 0.1$ で平均 0.08 と ϵ が小さくなるにつれ，有用性が下がった．DeSC データセットでは，$\epsilon = 1$ で順位相関が平均 -0.05，$\epsilon = 0.5$ で平均 0.08，$\epsilon = 0.1$ で平均 -0.05 と ϵ が小さくなるにつれ，有用性が下がるわけではなかった．



(a) NHANES データセット　　**(b)** DeSC データセット

図 5.2: 提案有用性 (オッズ比の順位相関)

提案方式ではオッズ比の MAE と順位相関が共に改善されていたため，提案方式で合成されたデータの方が元データと同じようなロジスティック回帰の結果を保持していると言える．すなわち，提案方式の合成データはより正しく病気の原因を分析することができる．

表 5.7 に，実際の例として NHANES データセットにおいて，元データで糖尿病に対する各因子 (カテゴリ属性) の影響をオッズ比を用いて調査した結果と，PrivBayes で作られた合成データで調査した結果，PrivBayes+提案手法で作られた合成データを用いて調査した結果の比較を示す．

オッズ比の MAE が 0.5 以上で極端に誤差が大きい箇所，影響度の順位が元データと変わった箇所，すなわち，正しく合成できていない箇所を太字で示した．オッズ比の MAE が 0.5 以上であるのは PrivBayes で 6 箇所，提案手法で 1 箇所のみであった．順位が変わった箇所は，PrivBayes で 19 箇所，提案手法で 7 箇所であった．

性別の因子を見てみると，元データは男性のオッズ比が 1.50，PrivBayes は 0.89，提案手法は 1.39 という結果となっている．元データと提案手法の合成データでは，男性であると女性に比べ．糖尿病のリスクが高まるという分析結果である．一方，PrivBayes の合成データでは男性であると女性に比べ，糖尿病のリスクが下がるといった結果が逆転した解析となった．つまり，提案手法の合成データで分析した方が，性別が糖尿病に与える影響を正しく解析できていた．

METs の項目では，元データと提案手法の合成データでは METs，すなわち，運動強度が高くなるにつれ，オッズ比が下がっている．つまり，高い運動強度で運動するほど，糖尿病になりにくくなる

ことを示している．一方，PrivBayes の合成データでは METs が中程度でオッズ比が最も低く 0.90，METs が低いで最も高く 1.10 となっている．また，オッズ比が全て 1 に近いことから METs は糖尿病に影響をあまり及ぼさないことを示している．つまり，分析の際に PrivBayes の合成データを用いると，影響を与えている因子の特徴が無くなってしまっている．一方で，提案手法の合成データを用いると，METs が糖尿病に与える影響を正しく解析できていた．

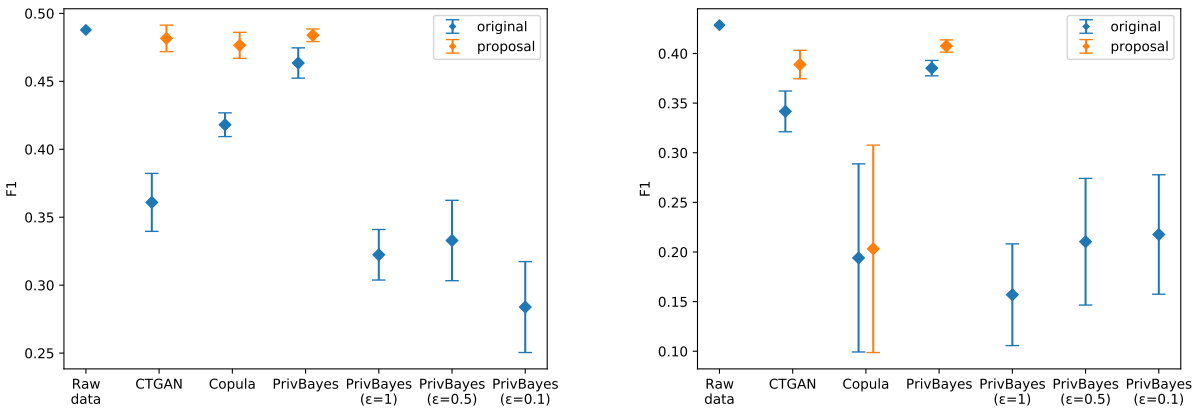
学歴の項目は，提案手法が最も正しく合成できていなかった箇所である．全ての順位が元データの真値とずれてしまっている．PrivBayes でも 5 箇所のうち 4 箇所の順位が変動していた．

5.2.3 実験 2: 機械学習モデルの有用性

図 5.3 に Random Forest の F 値を示す．NHANES データセット，ヘルスケアデータセットのどちらにおいても CTGAN，Gaussian copula，PrivBayes の全ての合成アルゴリズムで，提案手法の方が F 値が高くなった．

NHANES データセットでは CTGAN で平均 0.12，Gaussian copula で平均 0.06，PrivBayes で平均 0.02 ほど F 値が高くなっていた．元データ (Rawdata) で作成した Random Forest と比較すると，F 値が CTGAN では平均 −0.01，Gaussian copula では平均 −0.02，PrivBayes では平均 −0.01 とわずかに精度が落ちただけで，ほとんど元データと同じ精度でモデルを作成できていた．

DeSC データセットでは CTGAN で平均 0.05，Gaussian copula で平均 0.01，PrivBayes で平均 0.02 ほど F 値が高くなっていた．元データ (Rawdata) で作成した Random Forest と比較すると，F 値が CTGAN では平均 −0.03，Gaussian copula では平均 −0.22，PrivBayes では平均 −0.01 低くなっていた．CTGAN と PrivBayes はほとんど元データと精度は変わらなかったが，Gaussian copula はかなり精度が低かった．



(a) NHANES データセット　　**(b)** DeSC データセット

図 5.3: Random Forest の F 値

5.2.4 考察

既存合成アルゴリズム，提案手法，差分プライバシー合成アルゴリズムで作成された合成データでロジスティック回帰と機械学習モデルの有用性を評価し，比較を行った．

ロジスティック回帰については，提案手法が既存手法に比べて，大幅に有用性が改善された．そのため，提案手法は糖尿病 (目的変数) と因子 (説明変数) の関係をより正しく捉えてデータを合成できていたと言える．また，PrivBayes のみならず，CTGAN，Gaussian copula で説明変数を合成した場合にも提案手法は有効であり，有用性にも大きな差がなかった．説明変数を極端に作らない場合，例えば，元データには男性が 2000 人含まれているのに，合成したデータには男性が 10 人しか含まれていないなどの場合を除けば，説明変数の作り方によって有用性が大きく変化する可能性は低く，説明変数の合成方法によらずに安定して有用性の高いデータを合成できると考えられる．

機械学習モデル Random Forest の有用性についても，提案手法の方が既存手法に比べ，有用性が改善された．NHANES データセットでは PrivBayes，CTGAN，Gaussian copula で説明変数を合成した場合で，ほとんど有用性に差がなかった．しかし，DeSC データセットでは Gaussian copula で説明変数を合成した場合には，PrivBayes と CTGAN に比べ，有用性にかなり差があった．そのため，機械学習モデルの有用性は説明変数の合成手法やデータセットに影響を受ける考察される．

5.3 安全性評価実験

5.3.1 安全性の定義

本研究では，属性推定攻撃 [20] を用いて合成データの安全性を評価する．攻撃者は合成データと元データに含まれる特定の個人の情報を所持しており，それらを組み合わせてセンシティブな情報を推定することを仮定する．例えば，ここでは攻撃者が個人の性別，年齢，検診結果などの情報と元データから作られた合成データを使い，過去の既往歴などのセンシティブな属性を推定するような攻撃を想定している．この推定攻撃に対し，合成データがどれほど安全なのかを評価する．

5.3.2 実験方法

攻撃者が推定したいセンシティブな属性を NHANES データセットでは dep(鬱状態)，pir(貧困)，DeSC データセットでは問診の脳卒中の既往歴，心臓病の既往歴とする．各属性の情報を表 5.8 に示す．

攻撃者はこのセンシティブな属性を特定の個人のキー属性と合成データを用いて推定したいと想定する．所持するキー属性は，攻撃者の背景によるところがあり，どのキー属性を所持しているのかを想定するのが難しい．そのため，本研究はセンシティブ属性以外の全ての情報をキー属性として所持していると強い仮定をし，GCAP[13] で評価を行う．例えば，NHANES データセットは 10 属性の情報が含まれるが，攻撃者は，特定したいセンシティブ属性以外の 9 つの属性 (他のセンシティブ属性も含む) を全て所持していると仮定する．

各データセットでセンシティブ属性の分布に偏りがある．NHANES データセットでは dep=1 が 877 人，pir=1 が 884 人に対し，DeSC データセットでは脳卒中=1 が 111 人，心臓病=1 が 180 人で

ある．1 に該当する人がわずかであるため，全て 0 で推定すると，高い確率で推定できてしまう．そのため，本研究では NHANES データセットでは，各センシティブ属性で，1 である 500 人をランダムにサンプリング，0 である 500 人をランダムにサンプリングした 1,000 人を対象に推定し，評価を行う．DeSC データでも同様に 100 人ずつサンプリングを行い，200 人を対象に推定し，評価する．

全対象者 n 人で，i 番目のセンシティブ属性の真値を GCAP で推定した確率を $p_{gcap}^{(i)}$ とし，

$$A = \frac{1}{n}\sum_{i=1}^{n} p_{gcap}^{(i)} \tag{5.1}$$

で評価する．全ての対象者のセンシティブ属性を完全に誤って推定した場合は A は 0 になる．逆に，全ての対象者のセンシティブ属性を完璧に正しく推定できた場合は A は 1 になる．仮に $p_{gcap}^{(i)}$ が全て 0.5 で，ランダムに推定していたら，A は 0.5 になる．つまり，A=0 に近づくほど安全で，A=1 に近づくほど安全性が低い．A=0.5 だと，ランダムに推定しているのと変わらないという評価となる．

5.3.3 実験結果

NHANES データセットでの安全性評価を図 5.4 に示す．合成アルゴリズムとセンシティブ属性によって，提案手法の方が安全性が上がる場合と，下がる場合が見られた．

センシティブ属性 dep(鬱状態) では，提案手法の推定確率は CTGAN で平均 0.009 上がり，Gaussian copula で平均 −0.002 下がり，PrivBayes では平均 0.002 上がった．pir(貧困状態) では，提案手法の推定確率が CTGAN で平均 −0.002 下がり，Gaussian copula で平均 0.001 上がり，PrivBayes では平均 0.003 上がった．

PrivBayes の $\varepsilon = 1$，$\varepsilon = 0.5$，$\varepsilon = 0.1$ で差分プライバシーを満たす合成モデルでは，$\varepsilon = 1$ で推定確率が平均 0.506，$\varepsilon = 0.5$ で平均 0.509，$\varepsilon = 0.1$ で平均 0.511 と，ε が小さくなる につれ，推定確率が僅かに上がった．

センシティブ属性 pir(貧困状態) では，提案手法の推定確率は CTGAN で平均 −0.003 下がり，Gaussian copula で平均 0.001 上がり，PrivBayes では平均 0.002 上がった．pir(貧困状態) では，提案手法の推定確率が CTGAN で平均 −0.002 下がり，Gaussian copula で平均 0.001 上がり，PrivBayes では平均 0.003 上がった．

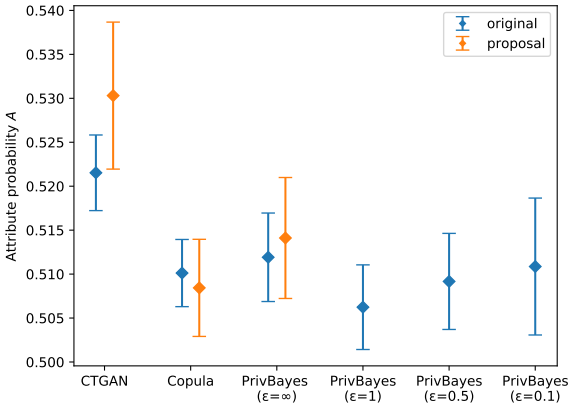
PrivBayes の差分プライバシーを満たす合成モデルでは，$\varepsilon = 1$ で推定確率が平均 0.520，$\varepsilon = 0.5$ で平均 0.517，$\varepsilon = 0.1$ で平均 0.508 と，ε が小さくなる につれ，推定確率が下がった．

DeSC データセットでの安全性評価を図 5.5 に示す．NHANES データセットと同様に，合成アルゴリズムとセンシティブ属性によって，提案手法の方が安全性が上がる場合と，下がる場合が見られた．

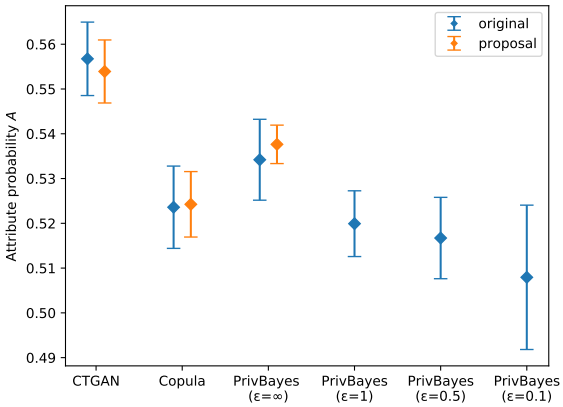
センシティブ属性の脳卒中では，提案手法の推定確率は CTGAN で平均 −0.005 下がり，Gaussian copula で平均 1.2×10^{-4} 上がり，PrivBayes では平均 3.8×10^{-5} 上がった．

PrivBayes の $\varepsilon = 1$，$\varepsilon = 0.5$，$\varepsilon = 0.1$ で差分プライバシーを満たす合成モデルでは，$\varepsilon = 1$ で推定確率が平均 0.502，$\varepsilon = 0.5$ で平均 0.523，$\varepsilon = 0.1$ で平均 0.513 と，ε =1 の時が最も推定確率が低かった．

センシティブ属性の心臓病では，提案手法の推定確率は CTGAN で平均 −0.007 下がり，Gaussian copula で平均 -8.5×10^{-4} 下がり，PrivBayes では平均 0.004 上がった．



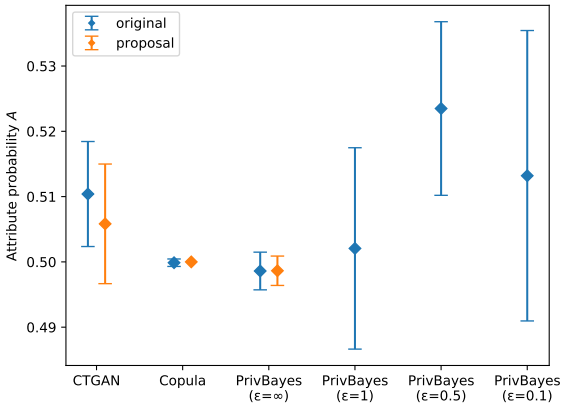
(a) dep



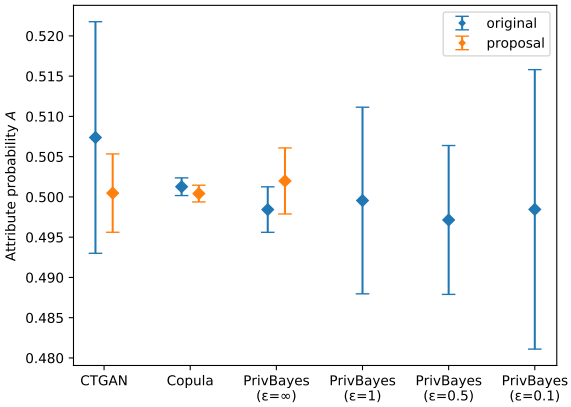
(b) pir

図 5.4: NHANES データセットにおける安全性評価

PrivBayes の差分プライバシーを満たす合成モデルでは，$\varepsilon = 1$ で推定確率が平均 0.499，$\varepsilon = 0.5$ で平均 0.498，$\varepsilon = 0.1$ で平均 0.498 と，ほとんど推定確率に差がなかった．



(a) 脳卒中



(b) 心臓病

図 5.5: ヘルスケアデータセットにおける安全性評価

5.3.4 考察

既存合成アルゴリズム，提案手法，差分プライバシー合成アルゴリズムで作成された合成データでセンシティブ属性の推定を行い，安全性を評価し，比較した．提案手法は既存手法と比べ，推定確率が -0.007 から 0.009 の範囲で変化した．ほぼ僅かな差であるため，提案手法と既存手法の属性推定による安全性はほとんど変わらないと言える．これは，提案手法と既存手法において，目的変数 1 つの合成手段が違うだけで，他の説明変数については，同じように合成しているためである．つまり，目的変数のみが大きく変わるだけで，他の説明変数にあまり差がないために属性推定のリスクが変わらないと考えられる．

表 5.5: 前処理後のデータセット項目

種別	項目	要素数	例
属性	年齢	連続値	42
属性	性別	2	男性
検診結果	bmi	連続値	21.5
検診結果	ldl コレステロール	連続値	127.0
検診結果	hdl コレステロール	連続値	58.0
検診結果	赤血球数	連続値	442.0
検診結果	中性脂肪	連続値	107.0
検診結果	収縮期血圧	連続値	121.0
検診結果	拡張期血圧	連続値	89.0
検診結果	got	連続値	26.0
検診結果	gpt	連続値	21.0
検診結果	ヘマトクリット値	連続値	46.5
検診結果	γ-gt	連続値	17.0
検診結果	血色素量	連続値	14.7
検診結果	血清尿酸値	連続値	6.5
検診結果	血清クレアチニン値	連続値	0.62
問診	「睡眠で休養が十分に取れている」	2	1: はい，2: いいえ
問診	「現在、たばこを習慣的に吸っている」	2	1: はい，2: いいえ
問診	「飲酒日の１日当たりの飲酒量清酒 １合（１８０ｍｌ）の目安」	4	1:1 合未満，2:1-2 合未満 3:2-3 合未満，4:3 合以上
問診	「日常生活において歩行又は同等の身体活動を 1 日 1 時間以上実施」	2	1: はい，2: いいえ
問診	「1 回 30 分以上の軽く汗をかく運動を 週 2 日以上、1 年以上実施」	2	1: はい，2: いいえ
問診	「夕食後に間食（3 食以外の夜食）を とることが週に 3 回以上ある」	2	1: はい，2: いいえ
問診	「20 歳の時の体重から 10kg 以上増加している」	2	1: はい，2: いいえ
問診	「この 1 年間で体重の増減が ± 3kg 以上あった」	2	1: はい，2: いいえ
問診	「現在、血圧を下げる薬を使用している」	2	1: はい，2: いいえ
問診	「現在、コレステロールを下げる薬を使用している」	2	1: はい，2: いいえ
問診	「医師から、慢性の腎不全にかかっているといわれたり、 治療 (人工透析) を受けたことがありますか」	2	1: はい，2: いいえ
問診	「医師から、脳卒中 (脳出血、脳梗塞等) にかかっている といわれたり、治療を受けたことがありますか」	2	1: はい，2: いいえ
問診	「医師から、心臓病 (狭心症、心筋梗塞等) にかかっている といわれたり、治療を受けたことがありますか」	2	1: はい，2: いいえ
問診	「医師から、貧血といわれたことがある」	2	1: はい，2: いいえ
病気	健診から三年後以内に糖尿病の診断記録がある	2	1: 糖尿病，0: 糖尿病ではない

表 5.6: 合成アルゴリズムの比較

	従来手法			提案手法
説明変数の合成 目的変数の合成	PrivBayes[5]	CTGAN[3]	Gaussian copula[4]	PrivBayes [5] CTGAN[3] Gaussian copula[4] + ロジスティック回帰
差分プライバシー	$\epsilon = \infty$(なし), $\epsilon = 1$ $\epsilon = 0.5$, $\epsilon = 0.1$	なし	なし	なし
合成ベース	BN	深層学習	copula	BN 深層学習 copula + 一般化線形モデル

表 5.7: PrivBayes と提案手法の OR の比較

	項目	元データ (真値)		PrivBayes		提案手法	
		オッズ比	影響度の順位	オッズ比	影響度の順位	オッズ比	影響度の順位
性別	男性	1.50	2	**0.89**	**1**	1.39	2
	女性	1	1	1	**2**	1	1
METs	高い	0.66	4	0.93	**2**	0.59	4
	中程度	0.71	3	0.90	**1**	0.67	3
	低い	0.78	2	1.10	**4**	0.72	2
	かなり低い	1	1	1	**3**	1	1
人種	Black	1	3	1	**2**	1	**4**
	White	0.43	1	**0.99**	1	0.50	1
	Hispanic	0.80	2	**1.37**	**4**	0.76	2
	Mexican	1.06	5	1.21	**3**	0.95	**3**
	Other	1.04	4	1.42	**5**	1.03	**5**
婚姻状態	未婚	1.26	4	**0.72**	**2**	1.48	4
	同棲中	1.36	6	**0.74**	**3**	**2.23**	6
	既婚	1.35	5	**0.84**	**4**	1.50	5
	別居中	1.15	3	1.15	**6**	1.26	3
	離婚	1	2	1	**5**	1	2
	未亡人	0.87	1	0.63	1	0.94	1
学歴	中卒	0.90	2	0.96	2	0.93	**1**
	高校中退	1	5	1	**3**	1	**4**
	高卒	0.91	3	0.82	**1**	1.11	**5**
	大卒	0.95	4	1.05	**5**	0.96	**2**
	大卒以上	0.85	1	1.01	**4**	0.89	**3**
鬱状態	鬱状態である	1.49	2	1.05	2	1.56	2
	鬱状態でない	1	1	1	1	1	1
貧困状態	貧困である	1.20	2	1.12	2	1.12	2
	貧困でない	1	1	1	1	1	1

表 5.8: センシティブ属性について

センシティブ属性	項目	属性値	人数
dep(鬱状態)	2 週間以内に気分が落ち込んだことがあるか	1(ある)	877
		0(ない)	3313
pir(貧困)	世帯年収が貧困レベルであるか	1(該当)	884
		0(該当しない)	3306
脳卒中 (既往歴)	「医師から、脳卒中 (脳出血、脳梗塞等) にかかっている といわれたり、 治療を受けたことがありますか」	1(はい)	111
		0(いいえ)	4889
心臓病 (既往歴)	「医師から、心臓病 (狭心症、心筋梗塞等) にかかっている といわれたり、 治療を受けたことがありますか」	1(はい)	180
		0(いいえ)	4820

第6章　まとめ

合成アルゴリズム PrivBayes で医療合成データを作成する際に，目的変数を指定してデータを合成することができないため，出力される合成データの特性がランダムに変わってしまう課題があった．これに対し，目的変数をロジスティック回帰モデルで合成する手法を提案し，有用性と安全性を評価した．

ロジスティック回帰の有用性は，オッズ比の MAE が 0.10 – 0.12，順位相関が 0.29 – 0.62 ほど改善した．機械学習モデル Random Forest の有用性については，F 値が約 0.02 改善した．安全性については，PrivBayes に比べ，提案手法のセンシティブ属性の推定確率は最大で 0.004 ほど上がった．推定確率は僅かに上がったものの，ほとんど PrivBayes と安全性が変わらないと言える．そのため，提案手法は安全性を損ねず，有用な医療合成データを作成可能である．同様に，PrivBayes のみならず，CTGAN，Gaussian copula にも提案手法が有効であることを示した．

参考文献

[1] 個人情報保護委員会，“パーソナルデータの適正な利活用の在り方に関する実態調査（令和元年度）”，(https://www.ppc.go.jp/files/pdf/personal_date_cases2019.pdf，2023 年 7 月参照).

[2] L. Sweeny, “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), pp. 557-570, 2006.

[3] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni, “Modeling tabular data using conditional gan”, Advance in Neural Information Pro-cessing Systems, 32:7335-7345, 2019.

[4] Neha Patki, Roy wedge, and Kalyan Veeramachaneni, “The Synthetic data vault”, IEEE International Conference on Data Science and Advanced Analytics, pp. 339-410, 2016.

[5] Jun Zhang, Graham Cormode, Cecilia M. Procopic, Divesh Srivastava, and Xiaokui Xiao. “Privbayes: Private Data Release via Bayesian Networks ”, ACM Transactions on Database Systems, 42(4), 2017.

[6] Thresa Stadler, Bristena Oprisanu and Carmela Tron-coso, “Synthetic Data - Anonymisation Groundhog Day”, USENIX Security Symposium, pp. 1451-1468, 2022.

[7] C. Dwork. “Differential privacy”, In Proceedings of the 33rd international conference on Automata, Languages and Programming-volume Part II, pp. 1-12. Springer-Verlag, 2006.

[8] Pedro F. Saint-Maurice, et al. “Association of Daily Step Count and Step Intensity With Mortality Amoug US Adults”, JAMA, 323(12) pp. 1151-1160, 2020.

[9] デジタルクロス，“大阪府、健康支援アプリで生活習慣病の発症確率を予測”，(https://dcross.impress.co.jp/docs/usecase/003009.html，2023 年 12 月参照).

[10] リソウ，“医療ビッグデータ活用により機械学習の優位性を解明”，(https://resou.osaka-u.ac.jp/ja/research/2022/20221011_1，2023 年 12 月参照).

[11] 池上和輝，伊藤聡志，菊池浩明，“匿名加工情報の応用（2）：各種傷病を予測する健康診断モデル”，コンピュータセキュリティシンポジウム，pp. 1230-1237，2020.

[12] Jannifer Taub, Mark Elliot, Maria Pampaka, and Duncan Smith, “Differential Correct Attribution Probability for Synthetic Data: An Exploration”, Privacy in Statistical Databases: International Conference, pp. 122-137, 2018.

[13] Markus Hittmeir，Rudolf Mayer，and Andreas Ekelhart，“A Baseline for Attribute Disclosure Risk in Synthetic Data”，ACM Conference on Data and Application Security and Privacy ，pp．133-143，2020.

[14] 岡田 莉奈，正木 彰伍， 長谷川 聡，田中 哲士，“統計値を用いたプライバシ保護擬似データ生成手法”，コンピュータセキュリティシンポジウム，pp．1366-1372，2017.

[15] 三浦 尭之，紀伊 真昇， 芝原 俊樹，市川 敦謙，岩花 一輝，奥田 哲矢，山本 充子，千田 浩司，“ベイジアンネットワークによる合成データ生成時のランダム性が持つ差分プライバシー性の評価”，コンピュータセキュリティシンポジウム，pp．1389-1396，2023.

[16] 三浦 尭之，紀伊 真昇， 芝原 俊樹，市川 敦謙，山本 充子，矢内 直人，“合成データ生成の出力を評価するメンバーシップ推定攻撃フレームワーク”，コンピュータセキュリティシンポジウム，pp．448-455，2022.

[17] Shagufta Mehnaz，Sayanton V. Dibbo and Ehsanul Kabir，Ninghui Li and Elisa Bertino，“Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models”，USENIX Security Symposium，pp．4579-4596，2022.

[18] Md Sakib Nizam Khan，Niklas Reje and Sonja Buchegger，“Utility Assessment of Synthetic Data Generation Methods”，Privacy in Statistical Databases: International Conference，pp．250-265，2022.

[19] Claire Little，Mark Elliot and Richard Allmendinger，“Comparing the Utility and Disclosure Risk of Synthetic Data with Samples of Microdata”，Privacy in Statistical Databases: International Conference，pp．234-249，2022.

[20] B．Fung，K.Wang，Chen and P.S.Yu，“Privacy-preserving data publishing: A survey of recent developments”，ACM Computing Surveys，Vol．42，No．4，pp．1-53，2010.

謝辞

本論文は筆者が明治大学大学院先端数理科学研究科先端メディアサイエンス専攻博士前期課程に在学中の研究成果をまとめたものである．本研究を遂行するにあたり多くの方々から多大なる御指導と御援助を賜りました．

特に明治大学総合数理学部先端メディアサイエンス学科の菊池浩明教授には，学部 2 年から修士 2 年までの 5 年間，多くのご指導をいただきました．ゼミでのディスカッションは研究について深く考えさせられることが多く，難しいこともありましたが，おかげで様々な学びと成長の機会となりました．深く感謝申し上げます．

また，研究に関しての助言や，時に息抜きで他愛もない会話に付き合っていただいた明治大学菊池研究室の皆様に感謝いたします．

本論文で使用した匿名加工医療情報を提供していただき，貴重なデータに触れる機会を与えてくださった DeSC ヘルスケア株式会社様に深く感謝申し上げます．

最後に，前期博士課程に進学する機会を与えてくれただけでなく，困った時に相談に乗ってくれたり，支えとなってくれた家族に深く感謝申し上げます．