Vol.2022-CSEC-97 No.6
Vol.2022-IOT-57 No.6
2022/5/19

医療データへの合成データ生成技術適用に向けた一検討

三浦 尭之[1,a)]　紀伊 真昇[1]　市川 敦謙[1]　千田 浩司[2]　木村 映善[3]

概要：医療データは要配慮個人情報を含む場合があり，プライバシーの観点などからその扱いには十分な注意が必要である．機微なデータを扱うには多くの倫理審査や安全管理対策が要求され，迅速な研究遂行上のボトルネックとなっている．そのため，元データの統計量の特徴を保持する合成データを生成し，実データの代用とする手法が数多く提案されている．本研究では，実際の医療データを用いて各種合成データ生成手法の安全性および有効性に関する実験的評価を行う．具体的には，統計量ベースの手法，機械学習ベースの手法，および深層学習ベースの手法を実装し，各手法によって生成される合成データの有用性を比較する．特に安全性の指標として差分プライバシーに着目し，各手法を拡張して差分プライバシーを満たす合成データを生成可能とするとともに，その有用性について考察する．

1. はじめに

Real World Data(RWD) を活用した研究は，Random Control Trial(RCT) に比較して内的妥当性が低いという欠点を指摘されつつも，近年の医療ビッグデータの出現や統計手法の発展によって外的妥当性の高いエビデンスを提供できる臨床研究手法としての評価が高まりつつある [1][2]．また医療分野での Artificial Intelligence (AI) に関する研究も多くは RWD に依存している．しかし RWD は医療ビッグデータとして活用できる可能性を秘めつつも，個人情報，とりわけ要配慮個人情報 [3] が含まれる場合，臨床研究上の倫理審査手続きやデータ保護において多大な工数を必要とし，迅速に研究を進めることが困難な状態である．特に AI に関する取り組みにおける障害として上位にデータアクセスの困難性があげられている [4]．そのため，患者の同意なしに第三者にデータを提供したり研究利用の承認プロセスを軽減したりするために，特定個人の識別可能性のリスクを軽減する匿名加工技術が開発された．しかし匿名加工技術は，特定個人の識別可能性のリスクを低減させるためのデータ加工によってデータの品質が低下することから，臨床研究における統計解析に支障をきたす場合がある．

上記をふまえ本研究では，従来の匿名化手法ではなく合成データを生成する手法の有効性に関する検討を行う．合成データは統計的な特徴を再現するようにデータを統計的手法や機械学習的手法を用いて生成したものであり，特定個人にかかわるデータを直接用いて生成するものではない．すなわち従来の匿名加工情報と異なり，合成データの個々のレコードは特定個人のデータに直接由来しないため [5]，合成データにはプライバシーのリスクが相対的に少ないと考えられている．また合成データによる本来のデータと近い形態でのデータ流通性の向上は，医療情報の相互運用性の改善に寄与し，医療ビッグデータの環境改善に貢献することが期待されている [6]．しかし合成データの普及には，データ利用者によって利用価値があると思われるような高いデータの有用性を容易に生成できることに加え，匿名性やプライバシーの侵害リスクが十分低くなるようなデータ加工を提供する手法を実現する必要がある [7]．

1.1 従来の匿名化と合成データ生成

k-匿名化 [8] などの従来の匿名化手法は，扱うデータが高次元の場合（1 レコードあたりの属性情報が多い場合），データの品質が著しく低下してしまうため有用性に課題がある [9]．これに対して，類似の統計量を持つ合成データをランダムに生成する合成データ生成技術は，高次元データに対しても品質が低下しにくいことが期待されている．

これまで提案されている合成データ生成技術は，統計量ベース，機械学習ベース，および深層学習ベースの手法に大別できる．公的統計分野では，古くから統計量ベースの手法が研究されており，実用にも供されている [10]．医用画像分野では GAN ベースの深層学習を利用した手法が多数提案されている [11]．

一方，合成データの有用性は，公開データセットなどを用いて実験を行い，元データと合成データの類似性をもって評価されることが多い．実際の医療データを用いて有用性を評価した例は筆者らが知る限りまだ少ない．

本研究では，統計ベース，機械学習ベース，および深層学習ベースの合成データ生成手法を実装し，公開データ

[1] NTT 社会情報研究所, 〒 180–8585 東京都武蔵野市緑町 3–9–11
[2] 群馬大学情報学部, 〒 371–8510 群馬県前橋市荒牧町 4–2
[3] 愛媛大学医学部医療情報学講座, 〒 791–0295 愛媛県東温市志津川 454
[a)] takayuki.miura.br@hco.ntt.co.jp

表 1　Adult Dataset の属性と形式

	属性	形式
1	age（年齢）	整数値 [17, 90]
2	workclass（就業形態）	カテゴリー（8 値）
3	fnlwgt（全体の中に占める重み）	整数値 [13769, 1484705]
4	eduation（最終学歴）	カテゴリー（16 値）
5	marital-status（婚姻状態）	カテゴリー（7 値）
6	occupation（職種）	カテゴリー（14 値）
7	relationship（世帯主との関係）	カテゴリー（6 値）
8	race（人種）	カテゴリー（5 値）
9	sex（性別）	カテゴリー（2 値）
10	capital-gain（資本利得）	整数値 [0, 99999]
11	capital-loss（資本損失）	整数値 [0, 4356]
12	hours-per-week（労働時間）	整数値 [1, 99]
13	native-country（国籍）	カテゴリー（41 値）
14	>50K, <=50K	カテゴリー（2 値）

セットに加え愛媛大学医学部附属病院の DPC(Diagnosis Procedure Combination) 診療報酬請求データセットを用いて，各手法によって生成される合成データの比較評価を行う．特に安全性の指標として差分プライバシー（DP: Differential Privacy）[12], [13] に着目し，各手法を拡張して差分プライベートな合成データを生成可能とするとともに，その有用性について考察する．

2. 実験設定

本研究の実験の設定，各手法，および評価方法について紹介する．

2.1 目的・概要

本研究の目的は，公開データセットおよび実医療データセットに対して，各種合成データ生成アルゴリズムによって生成された合成データの品質の比較評価をすることである．合成データ生成の手法としては，統計量ベースは岡田らによる方式 [14]，機械学習ベースはベイジアンネットワークによる合成データ生成 [15]，深層学習の GAN ベースは CTGAN[16] を用いて比較を行う（2.4 項）．

一方，前記の合成データ生成の手法は，一般にプライバシーの理論的保証はなされていない．そこで各手法について，プライバシーの指標として近年注目されている差分プライバシーを適用し，差分プライバシーを満たす合成データを比較対象とした．

2.2 データセット

実験で用いるデータセットの紹介とその定式化を行う．

2.2.1 Adult Dataset

本研究では公開データセットとして Adult Dataset[17] を用いた．これはカテゴリー値，連続値からなる 14 種類の属性を用いて，目的変数である「年収が 5 万ドル以上か否か」を推論するデータセットである（表 1）．なお 5 番目の属性である education-num は education と同等の情報なので今回は削除した．

2.2.2 DPC データセット

本研究では愛媛大学医学部附属病院の DPC（Diagnosis Procedure Combination，診断群分類）のデータを格納したデータウェアハウスから 2010 年度から 2013 年度にかけての 4 年間の DPC 関連データのうち，入院時の経路，救急搬送，入院の当時の体重・身長，喫煙の有無，妊娠の有無，食事・行動・歩行の自立度，診断群分類，手術，副傷病のどの要素が入院日数の重要な因子であるかを分析することを想定したデータセットを用いた．データセットのうち，一つでも欠損値がある項目を保有するレコードは除外した．DPC は調査のために多数のカテゴリカルデータの提出を義務づけている．本稿ではデータ保護の観点から属性と形式（表 2）および「合成データと元データの各種統計量の差異」,「合成データと元データからそれぞれ得られる学習モデルにおける正答率の平均二乗誤差」の結果のみを記載する．

表 2　DPC Dataset の属性と形式

	属性	形式
1	性別	カテゴリー（2 値）
2	入院経路コード	カテゴリー（7 値）
3	緊急入院	カテゴリー（2 値）
4	入院日数	整数値
5	身長	連続値
6	体重	連続値
7	喫煙有無	カテゴリー（2 値）
8	妊娠有無	カテゴリー（2 値）
9	食事の自立	カテゴリー（4 値）
10	行動の自立	カテゴリー（4 値）
11	歩行の自立	カテゴリー（5 値）
12	診断群大分類	カテゴリー（18 値）
13	手術	カテゴリー（9 値）
14	手術サブ分類	カテゴリー（10 値）
15	副傷病	カテゴリー（3 値）

2.2.3 定式化

属性の集合を $A_1, \ldots, A_d$ とおく．ここで属性とは Adult Dataset でいう年齢や就業形態などのことであり，$A_1 = \{17, 18, \ldots, 90\}, A_2 = \{\text{Private, Self-emp-not-inc, } \ldots\}$ などといったものが該当する．このとき,，データセット内の個人は $r \in A_1 \times \cdots \times A_d$ と表現することができる．そして，この個人に対応するレコードを一行として，多くのレコードを行列の形でまとめたものをデータセット $D \in (A_1 \times \cdots \times A_d)^N$ とする[*1]．考察しているデータセットの取りうる可能性全体の集合を $\mathcal{D}$ と書くこととする．

2.3 合成データ生成技術と安全性

合成データ生成は，図 1 のように二段階に分解することができる．一段目の生成パラメータ抽出は，元データから統計量を計算する，あるいは学習によって機械学習パラメータを得る操作である．ここで，得られた統計量やパラメータを**生成パラメータ**と呼ぶこととする（生成パラメータの取りうる値の集合を $Param$ とおく）．この操作は「生成パラメータ抽出 $\text{EXT} : \mathcal{D} \to Param$」と表せる．

次に二段目として，その生成パラメータ $\theta \in Param$

[*1] データセットの正確な記述としては多重集合がふさわしいが，直積の元という形式でも本稿の議論上は問題ない．

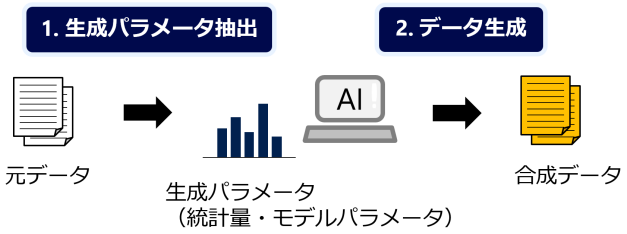

図 1 合成データ生成の処理手順

表 3 実験方式

	生成パラメータ	DP 実装
STAT[14]	統計量	diffprivlib[18]
BN[15]	グラフ構造・確率テーブル	PrivBayes[19]
CTGAN[16]	深層学習モデルパラメータ	Opacus[20]

を用いてランダムにデータを生成する「データ生成 $\mathrm{Gen}_\theta : \mathbb{R}^* \to \mathcal{D}$」がある．データ生成部では任意のレコード数のデータを生成することができるが，本研究では問題を単純化するため，入力データセットと同じ数のデータを出力することとする．

前述のとおり，本研究では合成データ生成の安全性指標として差分プライバシー（Differential Privacy, DP）を扱う．DP は多くの論文で論じられており，実装も充実している [18], [19], [20]．合成データ生成における DP は一般に，生成パラメータ抽出部の出力で保証する．これにより，生成パラメータから得られる合成データも DP を満たす．具体的には，関数 EXT を DP を満たすよう設計されたランダム化関数 $\hat{\mathrm{EXT}} : \mathcal{D} \to Param$ に置き換えることによって実現する．

実数 $\varepsilon, \delta > 0$ に対して，合成データ生成が (ε, δ)-差分プライバシーを満たすとは，任意の隣接データ $D \sim D'$ と任意のレンジ $S \subset Param$ に対して，次の不等式が成り立つことをいう：

$$\Pr[\hat{\mathrm{EXT}}(D) \in S] \leq e^{\varepsilon} \Pr[\hat{\mathrm{EXT}}(D') \in S] + \delta.$$

ここで，ε は 0 に近いほどプライバシー保護の度合が強いことを意味する（DP を考慮しない状態を $\varepsilon = \infty$ と解釈することもできる）．

抽出部のランダム性の持たせ方としては，生成パラメータが統計量の場合は Laplace メカニズムや Gaussian メカニズム，深層学習のモデルパラメータの場合は DP-SGD (Differentially-Private Stochastic Gradient Descent) [21] などで実現することが一般的である．DP を満たす合成データは一般に，保護度合いを強めるほど，すなわち ε を小さくするほど，品質は落ちることが知られている．

2.4 実験で用いる合成データ生成アルゴリズム

本研究で用いる 3 つの手法の詳細を紹介する（表 3）．実装は Python で行い，DP のフレームワークは Diffprivlib [18] や Opacus [20] を用いた．

2.4.1 STAT [14]

CSS2017 で岡田らは統計量に基づく合成データ生成技術を提案した（本稿では STAT と呼ぶ）[14]．匿名化の技術を競うコンテスト PWS Cup 2020[*2]および PWS Cup 2021[*3]で，ベースとなる手法としてサンプルコードが公開されている．

STAT の生成パラメータは，元データの属性ごとの度数分布の組 H と，各カテゴリー属性を二値ベクトル化（例えば属性 A の要素数を n としたとき，A の二値ベクトルは要素が一つだけ 1 で残りが 0 となる n 次元ベクトルとなる）し，結合して一つのベクトルとした際の各属性値の平均の組 μ および分散共分散行列 Σ からなる $\theta = (H, \mu, \Sigma)$ である．DP を満たす STAT（本稿では DP-STAT と呼ぶ）は，θ の各要素に適切なノイズを加えることによって実現可能である．本研究ではライブラリ Diffprivlib[18] のパッケージを用いて DP-STAT を実装した．

2.4.2 ベイジアンネットワーク [15]（BN）

ベイジアンネットワーク（BN） [15] は各属性間の因果関係を有向非巡回グラフで表現する確率モデルである．ここでノードは元データの各属性に対応し，エッジは因果関係の有無を表す有向エッジである．子ノードは隣接する親ノードの属性値を受けた条件付確率を学習し（テーブル状に保持），データ生成時は親ノード側から確率的に値を決定していく．

学習はデータからエッジの向きを学習する構造学習のステップと，各ノードの条件付確率の表を学習するステップがある．ここで生成パラメータは，エッジ情報（グラフ情報 G）と各ノードの確率表 T_p の値の組 $\theta = (G, T_p)$ である．DP を満たす BN（本稿では DP-BN と呼ぶ）としては，例えば PrivBayes [19] が提案されている．本研究では DataSynthesizer [*4] を利用して DP-BN を実装した．

2.4.3 CTGAN [16]

敵対的生成ネットワーク（GAN）は生成器 G と識別器 D という二つのニューラルネットワーク（NN）を競合的に学習させていくフレームワークである [22]．識別器 D は実際の訓練データと生成器 G が生成したデータを見分けるよう学習をする．生成器 G は識別器 D をだますよう学習を進める．両者を交互に学習させてき，学習終了時の生成器 G が出力するデータを合成データとして用いる．この場合の生成パラメータは生成器のモデルパラメータ θ_G である．

GAN ベースの合成データ生成技術として本研究では CTGAN [16] を採用した．CTGAN はテーブルデータに特化した GAN で，連続値からなる属性を混合ガウス分布で学習する．実装はその著者らによるもの[*5]を参考にした．DP を満たす CTGAN（本稿では DP-CTGAN と呼ぶ）は識別器 D の学習に DP-SGD を適用することで実現した．実装には Opacus [20] を用いた．

2.5 合成データの評価

本研究では，元データ D_{orig} と合成データ D_{syn} を次の観点で比較した．

[*2] https://www.iwsec.org/pws/2020/cup20.html
[*3] https://www.iwsec.org/pws/2021/cup21.html
[*4] https://github.com/DataResponsibly/DataSynthesizer
[*5] https://github.com/sdv-dev/CTGAN/tree/02f85b02

(1) 元データと合成データの統計的な分布の類似性
(2) 合成データを用いた機械学習の有用性
(3) DP を満たす合成データの品質

本研究では, Hellinger 距離, Kolmogorov–Smirnov 距離, および相関係数を用いて (1) を，2 値分類問題および回帰問題の機械学習モデルを用いて (2) を，そして各合成データ生成手法に DP を適用することで (3) を検討した.

2.5.1 Hellinger 距離

D_{orig} と D_{syn} の各属性の分布（ヒストグラム）を**Hellinger 距離**を用いて比較した.（離散確率分布の）Hellinger 距離とは，離散確率分布 $x = (x_i)_{i=1}^n, y = (y_i)_{i=1}^n$ に対して,

$$d(x,y) := \sqrt{\sum_{i=1}^{n}(\sqrt{x_i} - \sqrt{y_i})^2} \qquad (1)$$

で定まる距離である．Hellinger 距離は 0 から $\sqrt{2}$ の値をとる．二つの確率分布の Hellinger 距離が 0 に近いことはその確率分布が互いに近いことを意味する.

2.5.2 Kolmogorov–Smirnov 距離

D_{orig} と D_{syn} の各属性の分布（ヒストグラム）について，**Kolmogorov–Smirnov (KS) 距離**を用いて比較した．KS 距離は確率分布間の距離であり，KS 検定で用いられる統計検定量としてよく知られている．二つの確率分布の KS 距離が 0 に近いことはその確率分布が互いに近いことを意味する.

2.5.3 相関係数

D_{orig} と D_{syn} の整数値や連続値の数値属性同士の関係について，相関係数を用いて比較した．元データ D_{orig} と合成データ D_{syn} から相関係数の行列を計算し，両者の差の絶対値を可視化することで二つのデータを比較した．相関係数の差の絶対値が小さければ二つのデータが統計的に似通っていることを意味する.

2.5.4 機械学習モデル精度

合成データ D_{syn} を用いて学習された機械学習モデルを元データ D_{orig} を用いて評価する．精度が高ければ，合成データの分析の手法が元データの分析にも有用であること意味する．また，Adult Dataset では目的変数を「>50K, <=50K」の 2 値分類問題を，DPC データセットでは目的変数「入院日数」の回帰問題を解くこととした．分類問題は，XGBoost，決定木，サポートベクターマシン（SVM）を用いて行い，回帰問題は LightGMB を用いて学習させた．決定木と SVM は scikit-learn に含まれる実装を用い，決定木の深さは 5 以下とした。また，分類問題は正答率，回帰問題は予測誤差を平均二乗誤差（MSE）で評価した.

2.5.5 DP の適用

本実験では各合成データ生成手法に DP を適用して生成されるデータの品質がどのように低下するかを評価した．具体的な DP の適用方法は 2.4 項で紹介されているとおりである．実験では，DP を適用しない通常の合成データ生成（これを $\varepsilon = \infty$ の状態と記述する）に加えて，ε を $\varepsilon = 1, 2, 4, 8$ としたときの 4 通りの実験を行う．また，δ は多くの文献で用いられる $\delta = 10^{-5}$ を用いた.

3. 実験結果・考察

2 節で紹介した各手法で合成データを生成し，評価した結果を記す．各手法 5 回ずつ行い，その平均の値を記載しているが，DPC データセットに対する Bayesian Network による評価（図 4, 5，7，表 5）だけは，実験の都合から本稿では 1 回のみの結果を参考として記載していることに注意されたい.

3.1 Hellinger 距離，KS 距離

Adult Dataset における Hellinger 距離および KS 距離をそれぞれ図 2, 図 3 に, DPC データセットにおける Hellinger 距離および KS 距離をそれぞれ図 4，図 5 に示す．両データセットにおいて，各手法（DP-STAT は青，DP-BN はオレンジ，DP-CTGAN はグレーで表示）の Hellinger 距離および KS 距離を属性ごとに並べて表示し，同一色内の並びは左から DP 非適用（$\varepsilon = \infty$），$\varepsilon = 8, 4, 2, 1$ の結果を示している.

3.1.1 DP 非適用の場合

STAT, BN, CTGAN ともに，多くの属性において誤差は小さいことが確認できた．しかし hours-per-week, 入院日数属性の Hellinger 距離，および capital-gain, capital-loss, hours-per-week, 入院日数属性の KS 距離は例外的に誤差が大きくなっている.

hours-per-week 属性の Hellinger 距離については，連続値に対する Hellinger 距離の実装，および元データのグラフの形状が原因になっていると考えられる．連続値に対してはそのままでは Hellinger 距離を測ることはできないため，適当な区切り[*6]でヒストグラム化し，合計が 1 になるよう正規化し，式 (1) のとおりに計算している．このとき，同じ連続値の capital-gain, capital-loss 属性のように値域が広いものは，適当に区切りヒストグラム化しても一般に影響は少ないが，hours-per-week 属性のように値域は狭いが形状が単峰的で 1 か所に集中している場合は，区切りの影響が大きく，大きな誤差が生じ得ると考えられる．なお入院日数属性は非公開データのため，元データのグラフの形状などの考察はできないが，hours-per-week 属性と同様の傾向の可能性があると考えられる.

一方, Adult Dataset の capital-gain, capital-loss, hours-per-week 属性の KS 距離が大きいのは，元データのヒストグラムの形状が単峰的で 1 か所に集中しているためであると考えられる．KS 距離は累積密度関数の差の最大値に大きく影響されるため，そのような形状の分布は少しずれるだけで大きな誤差になると考えられる．こうした分布間距離の測り方による問題は今後の課題としたい.

3.1.2 DP 適用の場合

DP-STAT, DP-BN, DP-CTGAN ともに，ほぼ全ての属性において，DP 非適用の場合と比べ誤差が著しく大きくなった．DP を満たすため生成パラメータに故意のランダ

[*6] Python パッケージ numpy に実装されている関数 numpy.histogram_bin_edges などで自動的にとれる.

ムノイズを付加していることから，誤差が増加することは予想通りだが，実用の範囲とは言い難い．

DP-STAT は $\varepsilon = 8$ でも誤差が大きくなってしまい，DP-CTGAN は多くの属性において測定結果の分散が大きく，合成データの品質が安定しない結果になった．そして相対的には DP-BN の誤差が概ね小さいという結果を得た．

DP-STAT は度数分布にノイズを加えた上で白色化を行い，ノイズが付加された平均や分散共分散行列でデータを回転させていた．このノイズが度数分布に大きな変化を与えていると考えられる．すなわち，トータルのプライバシー予算 ε に対して，DP の合成定理 [23] に基づき度数分布および平均・分散共分散行列にそれぞれプライバシー予算 $\varepsilon/2$ を振り分けるが，元の属性数 d およびカテゴリー属性を二値ベクトル化した属性数 d^* に対して，分散共分散行列の $\frac{d^*(d^*+1)}{2}$ 個の要素それぞれが $\sum \varepsilon^* \leq \varepsilon/4$ となる ε^*–DP を満たす必要があるため，元の度数分布と大きく離れたものが作成されてしまうと考えられる．本稿の diffprivlib では，分散共分散行列の固有値・固有ベクトルにノイズを付加する手法となっているが，行列のサイズは，例えば Adult Dataset においては，$d^* = 128, d = 14$ より，128×128 であり，一つ一つの固有ベクトルに消費できる ε はトータルが $\varepsilon = 8$ だとしても $\varepsilon^* \leq 8/(4 \times 128) = 0.03125$ となり非常に小さい．

本研究で実装した，PrivBayes をベースとした DP-BN では，構造学習および確率表学習にそれぞれ $\varepsilon_1, \varepsilon_2$ のプライバシー予算を用いており，$\varepsilon = \varepsilon_1 + \varepsilon_2$ としている．構造学習では $d-1$ 回のノイズ付加を行う．すなわち各々のノイズ付加は $\frac{\varepsilon_1}{d-1}$–DP を満たすようにする．確率表学習においても同様に，d 回のノイズ付加を行い，各々は $\frac{\varepsilon_1}{d}$–DP を満たすようにする．ここで，DP-BN ではデータが二値ベクトル化されないため，DP-STAT と比べて個々成分にかかるノイズの実質のプライバシー予算が d^*/d 程度削減され，品質の低下が緩和されることが期待される．実際，Adult Dataset において，ヒストグラムの仕方に課題が残る連続値の fnlwgt, capital-gain, capital-loss 属性を除き，DP-STAT よりも誤差は小さくなっており，2 値属性などは誤差が非常に小さいことが分かる．しかしカテゴリー属性でも race, native-country 属性の誤差は $\varepsilon = 8$ でも無視できないほど大きく，さらなる分析や改善が必要と言える．

本研究で実装した DP-CTGAN は，SGD において各 Epoch であらかじめ設定した大きさのノイズ（本稿では $\mathcal{N}(0, I_d)$）を加えていき，合計で消費したプライバシー予算が ε を超えた段階で終了するという一般的な方式である．この方式で Adult Dataset などを学習させると，$\varepsilon = 1$ の場合，5 Epoch ほどで学習が停止していた．これは深層学習モデルの学習のためには不十分であり，また，生成モデルは一般に学習が安定しない傾向があるため，このような設定下では一層安定せず結果ごとの分散が大きい結果となった．Epoch 数を固定して合計が予め設定した ε になるようなサイズのノイズを付加する方式も実装可能であり今後の課題としたい．

3.2 相関係数

数値属性間の相関係数の誤差を評価した．Adult Dataset には，数値属性が age，fnlwgt，capital-gain，capital-loss の 5 種類あり，元データと合成データの相関係数の差のヒートマップは 5×5 で図 6 のとおりになった．濃い色が大きい値，薄い色が 0 に近い値を意味する．左から DP 非適用（$\varepsilon = \infty$），$\varepsilon = 8, 4, 2, 1$ としている．DPC データセットは数値属性が入院日数，身長，体重の 3 つのため，ヒートマップは 3×3 で図 7 のとおりになった．

3.2.1 DP 非適用の場合

図 6, 7 の左端のヒートマップから，特に STAT の誤差が小さいことが分かる．属性 A_i, A_j の相関係数 r_{ij} は，属性 A_i の分散 s_i, 属性 A_j の分散 s_j，および属性 A_i, A_j の共分散 s_{ij} から $r_{ij} = s_{ij}/\sqrt{s_i s_j}$ となるため，分散共分散の誤差が相関係数に影響を与える．しかし STAT は元のデータの分散共分散行列を保持する合成データを生成するため，相関係数も保持できる．ただし実際には，量子化による若干の誤差が生じている．

BN や CTGAN は基本的に，STAT と異なり基本統計量を保持するアルゴリズムではなく，STAT よりも同等以下の結果となっている．

3.2.2 DP 適用の場合

DP-STAT は分散共分散行列の各要素にノイズを付加する．3.1.2 節で述べたとおり，各要素のノイズは非常に小さなプライバシー予算から生成されるため，大きなノイズとなる．さらに r_{ij} の式より，相関係数はノイズが付加された分散の積に依存するため，ノイズの影響はより大きくとなると考えられる．

図 6 では，DP-STAT, DP-BN, DP-CTGAN は異なる傾向が見られるのに対し，図 7 では同様の誤差傾向が見られる．一方で 3.1.2 節でも触れたように DP-CTGAN は分散が大きく，現時点では有用性の評価が困難と言える．

3.3 機械学習モデル精度

Adult Dataset についての機械学習予測器の精度は表 4 のとおり，DPC データセットについての回帰の予測器の平均二乗誤差（MSE）は表 5 のとおりになった．表 5 のオリジナルと書いてあるのが元データで分析をした際の平均的な誤差の値であり，表の中の数値は各手法で生成された合成データで学習したモデルで，元データの入院日数を予測させたときの MSE の値である．

3.3.1 DP 非適用の場合

Adult Dataset については，各手法ともに DP 非適用の場合は 80%程度のテスト精度で，特に BN の精度がよい結果となった．DPC についても同様に良い結果で，オリジナルデータでの結果に近い MSE の値に近い結果になったが，これは，相関関係や因果関係の保持が今回の分析には重要であったためと考えられる．

3.3.2 DP 適用の場合

一方で，DP 適用の場合は Hellinger 距離などのほかの結果同様，ほどんどの場合，品質の劣化が確認された．しかし，Adult Dataet では DP-STAT，DP-CTGAN が大きく

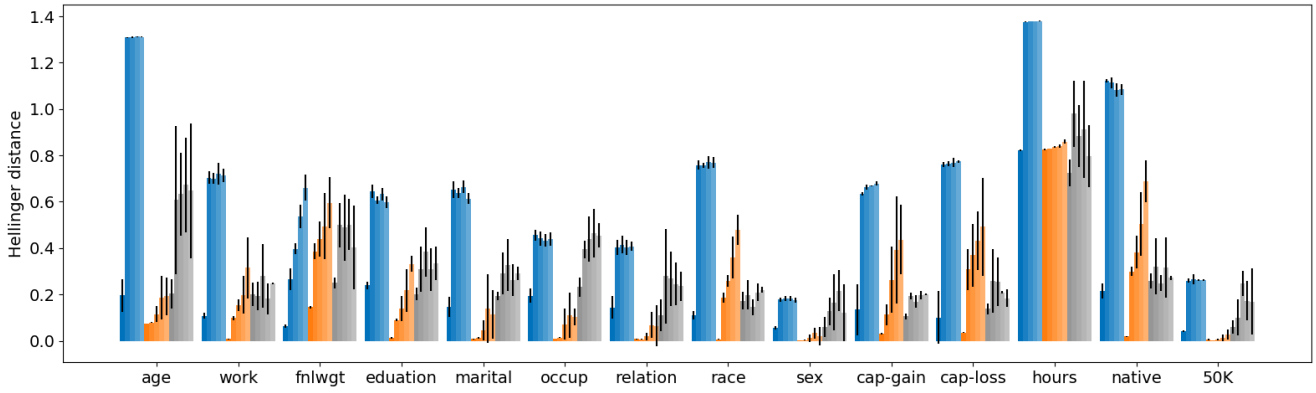

図 2 Hellinger 距離（Adult Dataset）：各属性左から DP-STAT（青），DP-BN（オレンジ），DP-CTGAN（グレー）．同一色内の並びは $\varepsilon = \infty, 8, 4, 2, 1$.

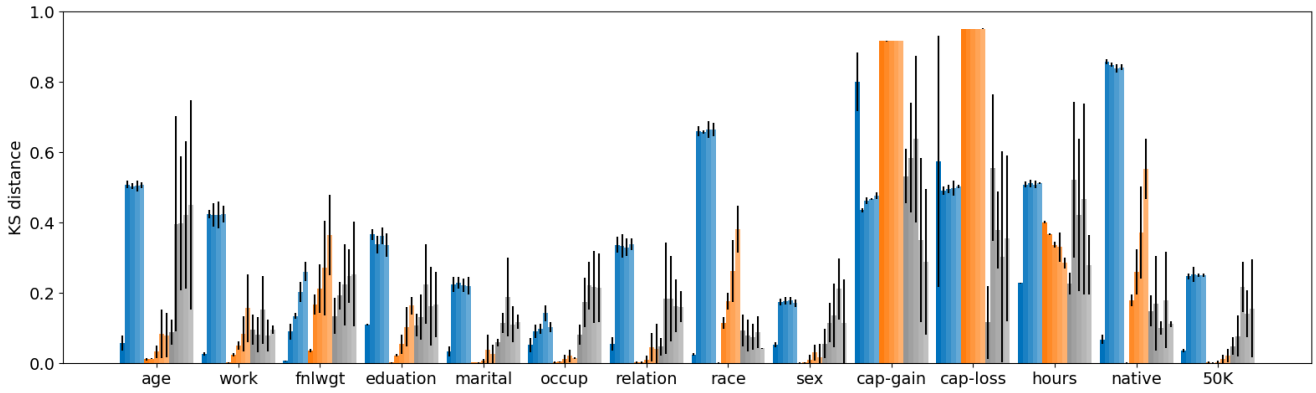

図 3 KS 距離（Adult Dataset）：各属性左から DP-STAT（青），DP-BN（オレンジ），DP-CTGAN（グレー）．同一色内の並びは左から $\varepsilon = \infty, 8, 4, 2, 1$.

表 4 機械学習モデル精度：分類精度（Adult Dataset）

モデル	ε	DP-STAT	DP-BN	DP-CTGAN
XGBoost	∞	0.79272	0.83664	0.81504
	8	0.43472	0.82972	0.73598
	4	0.52994	0.81386	0.52010
	2	0.49552	0.79634	0.60720
	1	0.46756	0.79742	0.61404
決定木	∞	0.77526	0.83394	0.80246
	8	0.41082	0.82704	0.73644
	4	0.51076	0.82002	0.55340
	2	0.43420	0.79954	0.65396
	1	0.40748	0.80796	0.59106
SVM	∞	0.78904	0.82484	0.81314
	8	0.41120	0.81678	0.73438
	4	0.55134	0.80564	0.57872
	2	0.48828	0.78548	0.60098
	1	0.49426	0.79012	0.58748

品質を損ねたのに対して，DP-BN は数ポイントの低下にとどまった．Bayesian Network は属性間の因果関係を学習し，その因果関係に基づいて学習を進める．ε が小さくなるにつれて，Hellinger 距離などの誤差が大きくなっている一方で，機械学習モデルの精度がよかったのは，目的変数にとって特に重要な因果関係を DP 適用下でも学習できていたため，そういった関係を保持した合成データが生成

表 5 LightGBM 機械学習モデル精度：MSE（DPC Dataset）

オリジナル	ε	DP-STAT	DP-BN（参考）	DP-CTGAN
17.4428	∞	18.999098	17.973505	21.605959
	8	22.878132	53.925526	32.363697
	4	35.354008	51.973053	21.811939
	2	52.176369	44.820789	21.049965
	1	61.736401	82.179337	40.861458

できていたからだと考えられる．

DPC では，DP-CTGAN が $\varepsilon = 4, 2$ のときは DP 非適用と同様の結果になったのは，生成モデルの学習が不安定であることが第一の原因として考えられる．加えて，Hellinger 距離などの結果のように，属性単体の再現はできていなくても，深層学習モデルが複数属性の関係を学習していて，特に入院日数のような数値属性とカテゴリ属性の間の複雑な関係性を再現できていた可能性も考えられる．こうした関係性をさらに可視化して考察していくことも今後の課題としたい．

4. まとめ

本研究では統計ベース，機械学習ベース，および深層学習ベースの合成データ生成手法に差分プライバシーを適用して実装し，公開データセットの Adult Dataset および愛

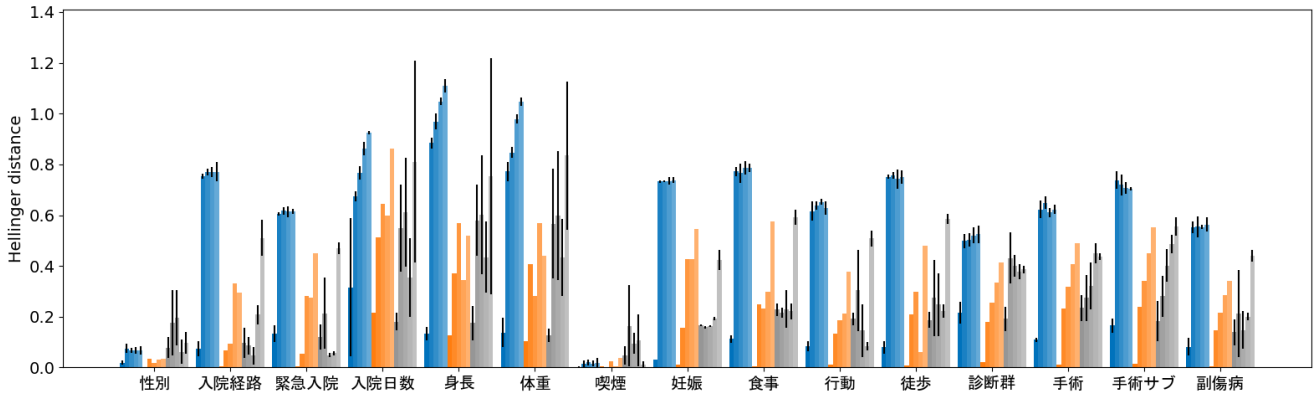

図 4 Hellinger 距離（DPC Dataset）：各属性左から DP-STAT（青），DP-BN（オレンジ），DP-CTGAN（グレー）．同一色内の並びは左から $\varepsilon = \infty, 8, 4, 2, 1$．DP-BN は 1 回のみの測定結果のため参考．

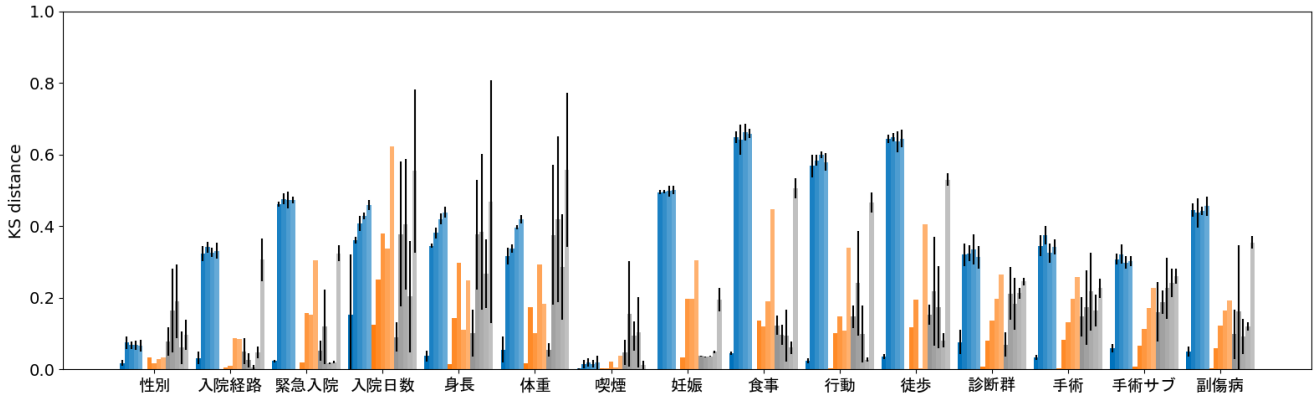

図 5 KS 距離（DPC Dataset）：各属性左から DP-STAT（青），DP-BN（オレンジ），DP-CTGAN（グレー）．同一色内の並びは左から $\varepsilon = \infty, 8, 4, 2, 1$．DP-BN は 1 回のみの測定結果のため参考．

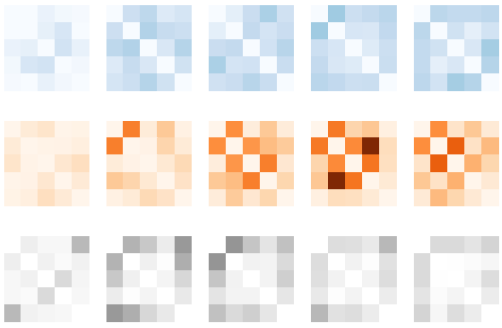

図 6 相関係数の差（Adult Dataset）：上段が DP-STAT（青），中段が DP-BN（オレンジ），下段が DP-CTGAN（グレー）．左から $\varepsilon = \infty, 8, 4, 2, 1$ の結果．数値属性の 5 属性を取り出したものの相関係数を配置．濃いほど大きい値．

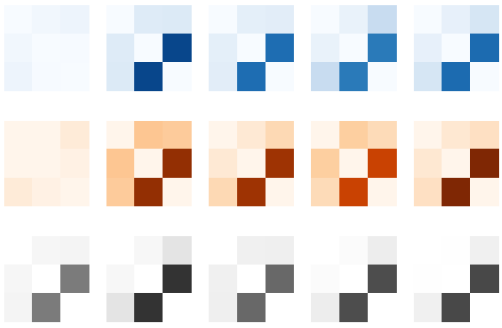

図 7 相関係数の差（DPC Dataset）：上段が DP-STAT（青），中段が DP-BN（オレンジ），下段が DP-CTGAN（グレー）．左から $\varepsilon = \infty, 8, 4, 2, 1$ の結果．数値属性の 3 属性を取り出したものの相関係数を配置．濃いほど大きい値．BN は 1 回のみの測定結果のため参考．

媛大学医学部附属病院の DPC データセットを用いて，各手法によって生成される合成データの比較評価を行った．結果として差分プライバシー非適用の場合は相対的に良い結果が得られた一方で，差分プライバシーを適用した場合は各手法のデータの品質が大きく落ちることが明らかになった．考えられる要因としては，いずれの手法も差分プライバシー化にのためを繰り返しノイズを付加するため，トータルのプライバシー予算 ε の値を大きくとっても，個々のプライバシー予算は小さくなり，ノイズが強くなってしま

うことが挙げられる．統計ベースの手法は個々のプライバシー予算が特に小さくなってしまうが，改善の余地があるため今後の課題としたい．また，深層学習ベースよりも機械学習ベースの有用性が高い場合が多く見られた点は興味深い．一方で，差分プライバシーを適用した深層学習ベースの合成データ生成手法は近年盛んに研究が進められているため，より高い品質の合成データ生成が期待できるアルゴリズムを検討し，実装評価を行う予定である．

別のアプローチとして，合成データ生成アルゴリズムに内在するランダム性を利用して，安全性を理論的に保証する研究も行われている [24], [25]．すなわち，差分プライバシーやその他の安全性指標を満たすためのノイズの強度をよりタイトに設定できる可能性がある．今後はこのようなアプローチも含め，有用性と安全性を高いレベルで両立させた合成データ生成手法の確立を目指す．

倫理に関する配慮事項

本研究は，「人を対象とする医学研究に関する倫理指針」に基づき，愛媛大学医学部の倫理審査委員会の承認を得て実施した．研究課題名「統計的特徴を維持した合成データ生成手法の品質評価」（承認番号 2012001）

参考文献

[1] Zahra Azizi, Chaoyi Zheng, Lucy Mosquera, Louise Pilote, and Khaled El Emam. Can synthetic data be a proxy for real clinical trial data? a validation study. *BMJ open*, 11(4):e043497, 2021.

[2] Anat Reiner Benaim, Ronit Almog, Yuri Gorelik, Irit Hochberg, Laila Nassar, Tanya Mashiach, Mogher Khamaisi, Yael Lurie, Zaher S Azzam, Johad Khoury, et al. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR medical informatics*, 8(2):e16492, 2020.

[3] 個人情報保護委員会. 個人情報の保護に関する法律についてのガイドライン（通則編）. `https://www.ppc.go.jp/personalinfo/legal/2009_guidelines_tsusoku/`.

[4] Jacques Bughin, Eric Hazan, Sree Ramaswamy, Michael Chui, Tera Allas, Peter Dahlstrom, Nicolaus Henke, and Monica Trench. Artificial intelligence: the next digital frontier? 2017.

[5] Jingchen Hu. Bayesian estimation of attribute and identification disclosure risks in synthetic data. *arXiv preprint arXiv:1804.02784*, 2018.

[6] Khaled El Emam and Richard Hoptroff. The synthetic data paradigm for using and sharing data. *Cutter Executive Update*, 19(6), 2019.

[7] Jerome P Reiter. New approaches to data dissemination: A glimpse into the future (?). *Chance*, 17(3):11–15, 2004.

[8] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[9] Charu C Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, volume 5, pages 901–909, 2005.

[10] 独立行政法人 統計センター. 教育用擬似ミクロデータの開発とその利用〜非衛生 16 年全国消費実態調査を例として〜. `https://www.nstac.go.jp/services/pdf/sankousiryou2407.pdf`, 2012.

[11] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[12] Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006.

[13] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[14] 岡田莉奈, 正木彰伍, 長谷川聡, and 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. In **コンピュータセキュリティシンポジウム** *2017* **論文集**, volume 2017, oct 2017.

[15] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[16] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32:7335–7345, 2019.

[17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[18] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. Diffprivlib: the IBM differential privacy library. *ArXiv e-prints*, 1907.02444 [cs.CR], July 2019.

[19] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4), October 2017.

[20] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

[21] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[23] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[24] 三浦尭之, 紀伊真昇, 芝原俊樹, 市川敦謙, and 千田浩司. 合成データ生成のランダム性に内在する安全性の評価. In **コンピュータセキュリティシンポジウム** *2021* **論文集**, pages 268–275, oct 2021.

[25] Zinan Lin, Vyas Sekar, and Giulia Fanti. On the privacy properties of gan-generated samples. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1522–1530. PMLR, 13–15 Apr 2021.