

医療ビッグデータを用いた AI による合成医療データの生成 および品質評価手法の開発

日本大学 理工学部 応用情報工学科
関 弘翔

1. はじめに

臨床試験をはじめとする医薬品開発では、試験条件に合致する被験者の選定に難渋するケースが多い。そのため、新薬開発等に必要となる被験者数を確保するのに膨大な時間や資金が必要となり課題とされている。臨床試験シミュレーション (clinical trial simulation; CTS) は、医薬品開発における意思決定のための重要な手段である。コンピュータ上で臨床試験をシミュレーションすることで、より効果的かつ効率的な臨床試験をデザインすることが CTS の目的である。現実的なシナリオを得るためには、シミュレーションに用いる仮想被験者が予定している臨床試験対象の被験者属性を適切に再現している必要がある¹⁾。仮想被験者を生成する一般的な方法は、ブートストラップリサンプリング (bootstrap resampling) または多変量正規分布 (multivariate normal distributions) である²⁾。ブートストラップ法は、過去の同種の臨床試験データなど、実際のデータから復元抽出を行うため「性別」「年齢」「体重」「検査値」といったデータが持つ属性の間の相関構造を維持できるが、過去に観測されていない母集団を推定できない問題がある。多変量正規分布は、分布を仮定することで観測していない属性を持った仮想被験者も新規に生成できるが、属性同士が線形に関係することに加え単峰性の分布を持つという、比較的強力な仮定が必要となる。これにより現実的な属性間の相関構造等を表現しきれず、現実離れした属性を持つ被験者を生成し得る問題がある。

また、現実の医療を反映した実臨床現場から得られるリアルワールドデータ (real world data; RWD) の利活用が、臨床試験などの医薬品開発や臨床研究領域において期待されている。RWD の活用は、質の高い観察研究の実現や希少疾患の患者数確保など様々な恩恵をもたらすことから国際的に注目されている。しかし、秘匿性の高さから複数組織間での共有の障壁は高く、RWD の価値を最大限活かしているとは言えない現状がある。

本研究の目的は、発展著しい人工知能 (artificial intelligence; AI) 技術、ディープラーニング (deep learning) を応用し、RWD を基にした表形式合成医療データの生成を行うことで、先述の問題の解決を図ることである。合成医療データ生成 AI により、従来よりも現実らしい仮想被験者生成が実現できるとともに、生成したデータは非実在データであるため、プライバシー情報開示のリスクを最小限に抑え、オープンな医療ビッグデータの実現に寄与できる。

本研究のように表形式データを AI で合成する研究例は多くある^{3,4)}。例えば 2021 年、米国の国立衛生研究所 (NIH) が Syntegra 社の AI により COVID-19 患者記録の合成データを生成している。しかし、既報における合成医療データの品質評価には、単一属性の分布や一対相関が用いられている。これらの手法は、年齢など属性値の分布や相関が本物と類似するかの評価にとどまり、属性の組み合わせが患者として妥当かを明確に評価するには至っていない。そこで、本研究では、合成医療データ生成手法および、その評価方法の確立に大規模な RWD を用いて取り組み実現することで、従来は難しかった、合成データの属性などの組み合わせが現実的であるか否かの評価をする手法の実現を目指す。

2. 方法

本研究は日本大学薬学部臨床研究に関する倫理審査委員会の承認(20-012)を得て実施した。AI、特にディープラーニング技術を用いた生成 AI (生成モデル) を医療ビッグデータである RWD で学習し、合成医療データを生成した。本研究で使用した RWD は、「慢性腎臓病」「認知症」「2 型糖尿病」「肺がん」「リウマチ」「移植患者」の計 6 つの疾患等からなる総計約 300 万人規模の患者データである。これらの RWD は、メディカル・データ・ビジョン (株) より購入し、許諾を受けて使用した。生成モデルには、表形式データに対応した敵対的生成ネットワーク (generative adversarial networks; GAN) の CTGAN³⁾ や、拡散モデル (diffusion model) の TabDDPM⁴⁾を採用した。さらに表形式データに対応した変分オートエンコーダ (variational autoencoder; VAE) である TVAЕ³⁾も比較に用いた。

また、合成データの評価方法に関し、属性の組み合わせが医療データ (患者データ) として妥当であるかに明確な尺度を設けることは困難である。そこで合成データの属性の組み合わせを評価する方法として、異常検知手法を応用することを提案した。組み合わせが妥当か否かは、実在する患者らしいかどうかと言える。よって、実在の患者データを用いて異常検知 AI を学習し、異常検知 AI が出力する異常度を見ることで、実在する患者らしいかを評価できると考えた。また、表形式データをそのまま入力として異常検知 AI を学習した場合、属性間の関係に着目できない可能性があるため、表形式データの特徴抽出を行う AI も組み合わせて評価する手法を提案した。さらに、合成データが何を根拠に異常と判断されたかを、説明可能な AI (explainable AI; XAI) により可視化する手法も取り入れた。

本研究では、ディープラーニングに決定木の考えを取り入れることで表形式データの解釈可能な特徴表現を得る TabNet⁵⁾と、異常検知手法であるアイソレーションフォレスト (isolation forest)⁶⁾を組み合わせで大規模 RWD を学習し、表形式医療データとしての品質を評価可能な AI を構築した。XAI には協力ゲーム理論の Shapley 値を機械学習に応用した SHAP (Shapley additive explanation)⁷⁾を採用した。図 1 に、本研究で提案する表形式合成データの品質評価 AI の概要を示す。

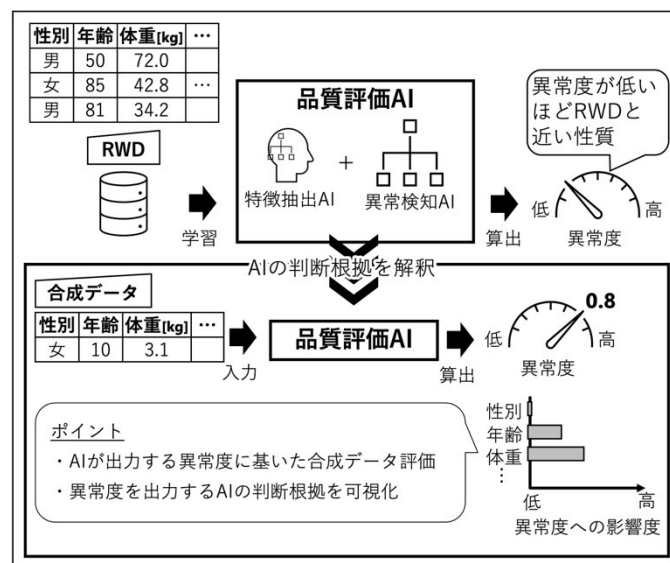


図 1. 提案する表形式合成データの品質評価 AI の概要

3. 合成患者データの生成

本研究では、薬物動態解析を行う研究⁸⁾への応用も見据え、生成する患者属性として、性別（SEX）、年齢（AGE）[year]、体重（TBW）[kg]、身長（HT）[cm]、臨床検査値として肝機能値の ALT[U/L]、AST[U/L]および、腎機能値の SCR[mg/dL]とした．表 1 に使用したデータを示す。ここで使用可能データとは、本研究で採用した患者属性を全て有するデータ数のことである。さらに、使用データを学習用、検証用、テスト用で 8:1:1 に分割して用いた。表 2 に学習データの基本統計量を示す。

表 1. 疾患別使用症例数

種類	n（対象患者数）[人]	使用可能データ数	使用データ数
慢性腎臓病	695,424	555,630	1000
認知症	917,868	537,644	1000
2 型糖尿病	354,001	1,048,576	1000
肺がん	451,027	353,411	1000
リウマチ	370,527	163,961	1000
移植患者	165,268	168,648	1000

表 2. 学習データ（4800 件）の基本統計量

	AGE[year]	TBW[kg]	HT[cm]	ALT[U/L]	AST[U/L]	SCR[mg/dL]
平均	74.3	55.8	156.9	25.8	33.4	1.39
標準偏差	12.9	13.3	10.2	51.6	87.1	1.78
最小値	1.0	9.7	50.0	1.0	3.0	0.19
中央値	76.0	54.5	157.0	17.0	22.0	0.83
最大値	100.0	164.1	191.0	1866.0	3139.0	20.74

表 2 のデータを用いて、表形式データ用の生成モデルである TVAE、CTGAN、TabDDPM の学習を行った。TVAE、CTGAN、TabDDPM の実装には SynMeter⁹⁾を応用した。表 3~5 に、各モデルでテストデータと同じ件数分生成した合成患者データの基本統計量を示す。また、図 2 に学習データの相関と合成データの相関の差分の絶対値を可視化したものを示す。

表 3. TVAE（600 件）の基本統計量

	AGE[year]	TBW[kg]	HT[cm]	ALT[U/L]	AST[U/L]	SCR[mg/dL]
平均	79.1	54.0	156.9	16.7	20.3	0.73
標準偏差	8.37	9.4	7.5	6.1	4.5	0.38
最小値	41.0	29.3	132.0	2.0	10.0	0.37
中央値	80.0	52.2	157.0	15.0	20.0	0.80
最大値	95.0	83.2	176.0	37.0	42.0	8.22

表 4. CTGAN（600 件）の基本統計量

	AGE[year]	TBW[kg]	HT[cm]	ALT[U/L]	AST[U/L]	SCR[mg/dL]
平均	70.6	47.3	154.3	20.8	43.9	1.45
標準偏差	15.8	12.3	11.0	30.9	55.7	1.79
最小値	2.0	-5.2	119.0	-5.0	-4.0	0.14
中央値	72.0	47.0	154.0	17.0	36.0	1.02
最大値	105.0	89.2	192.0	590.0	758.0	14.90

表 5. TabDDPM（600 件）の基本統計量

	AGE[year]	TBW[kg]	HT[cm]	ALT[U/L]	AST[U/L]	SCR[mg/dL]
平均	73.5	59.3	153.8	78.2	158.3	1.73
標準偏差	17.6	27.9	23.5	323.9	626.8	3.73
最小値	1.0	9.7	50.0	1.0	3.0	0.19
中央値	76.0	54.5	157.0	17.0	22.0	0.83
最大値	100.0	164.1	191.0	1866.0	3139.0	20.74

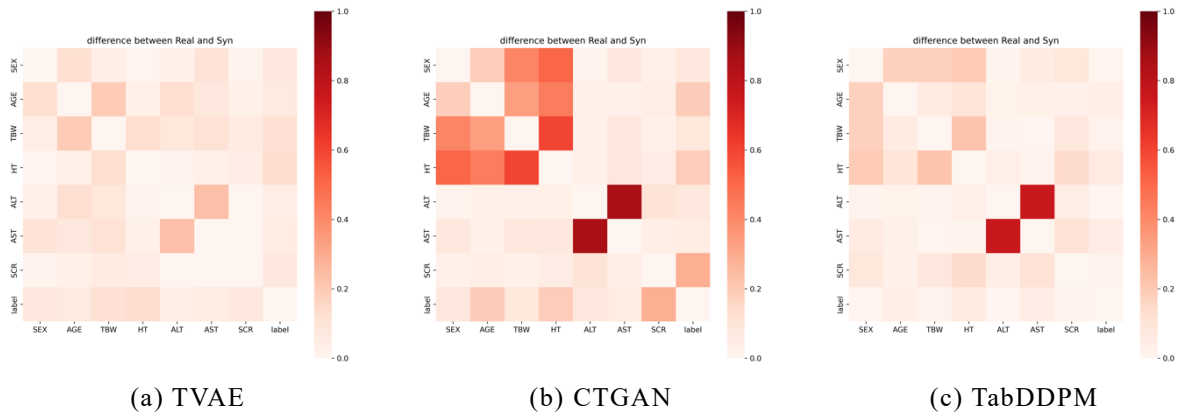


図 2. 学習データと合成データの属性間相関の差分

VAE は、平均に近いデータを生成しやすくなる特徴があり、表 2 と表 3 を比較すると、生成できている数値の範囲が学習データよりも狭くなっていることが確認できる。一方で図 2(a) の相関は学習データとよく類似しており、ある範囲においては品質が高いデータが生成されていることが示唆された。CTGAN は、連続値に多峰性分布を仮定してそれぞれのモードとモード内での値を決定するモード別正規化を取り入れている。表 2 と表 3 を比較すると、CTGAN は学習データが持つ値域を TVAE よりも再現できたが、それと引き換えに現実にはあり得ない負の値を生成することがあった。また、図 2(b) の相関は学習データと乖離する属性が多く、組み合わせが妥当でない合成データの生成が示唆された。TabDDPM は四分位に基づく正規化を取り入れることで学習データの最小値や最大値付近のデータにも上手く対応できており、表 2 と表 4 を比較すると学習データが持つ値域を最もよく再現していた。また、図 2(c) の相関は肝機能値を表す ALT と AST の間の相関を除けば、学習データの性質をよく再現していた。ALT と AST の相関の再現が他の属性に比べて劣った理由は、表 2 に示す通り、値域が広いことに起因している。

4. 合成患者データの評価

表 2 に示した統計量を持つ学習データを用いて異常検知 AI であるアイソレーションフォレストを学習した。学習したアイソレーションフォレストに、表 3~5 に示した統計量を持つ合成患者データを入力し、異常検知を行った。アイソレーションフォレストの実装には PyOD¹⁰⁾を応用した。次に、表 2 に示した統計量を持つ学習用の患者データを用いて表形式データの特徴抽出を行う TabNet を学習した。学習した TabNet に、同学習データを入力して特徴抽出を行い、抽出した特徴を用いてアイソレーションフォレストの学習を行った。学習した TabNet に表 3~5 に示した統計量を持つ合成患者データを入力して特徴抽出を行い、学習したアイソレーションに特徴抽出した合成患者データを入力して異常検知を行った。TabNet の実装には PyTorch Tabular を応用した¹¹⁾。さらに、学習したアイソレーションフォレストに XAI の SHAP を適用し、判断根拠の可視化を行なった。表 6 に、各方法で異常度（0.0~1.0 としている）が 0.6 以上と判断された合成データの件数および平均値を示す。

表 6. 各手法で異常度が 0.6 以上となったデータの平均値

手法	件数	AGE [year]	TBW [kg]	HT [cm]	ALT [U/L]	AST [U/L]	SCR [mg/dL]
TVAE data	0	-	-	-	-	-	-
TVAE data +TabNet	7	93.1	47.2	149.6	17.6	20.3	0.74
CTGAN data	11	58.9	32.4	145.7	13.8	71.9	2.75
CTGAN data +TabNet	13	85.2	36.7	155.5	20.0	39.4	4.23
TabDDPM data	15	54.8	104.4	151.6	270.0	642.5	4.96
TabDDPM data + TabNet	47	78.1	66.2	147.5	56.1	87.8	8.30

表 6 より、TVAE が最も異常と判定される件数が少なかった。異常検知が、学習データからの逸脱具合を測るものであるため、平均的な性質を持つデータを生成する TVAE は異常と判定されにくかったものと考えられる。また、TabNet で特徴抽出を行うことで、異常度が高くなるデータがいくつか存在した。これらすべては 91~94 歳のデータになっている（標準偏差 1.2 歳である）ことが確かめられた。図 3 に、SHAP により可視化した TVAE の合成データに対する異常検知 AI の判断根拠の外観を示す。各属性の値（Feature value）をカラースケールで示しており、横軸が予測への寄与度を示す SHAP 値を示している。SHAP 値が大きい正の値であるほど、異常と判断されるのに寄与していることを意味する。図の縦軸の属性は、上から平均的な SHAP 値の絶対値が大きい順、すなわち、重視されやすい特徴順に並んでいる。図 3 より、特徴抽出前後で重要視される特徴が大きく変わっている。特徴抽出後は年齢、体重、SCR の順で重視されていることが確認できたため、異常と判断されたデータの中でそれらの統計量を確認したところ、SCR の標準偏差も 0.07 mg/dL と小さかった。よって、ある特定の組み合わせを持つデータの異常度を高く算出していることが確認できた。

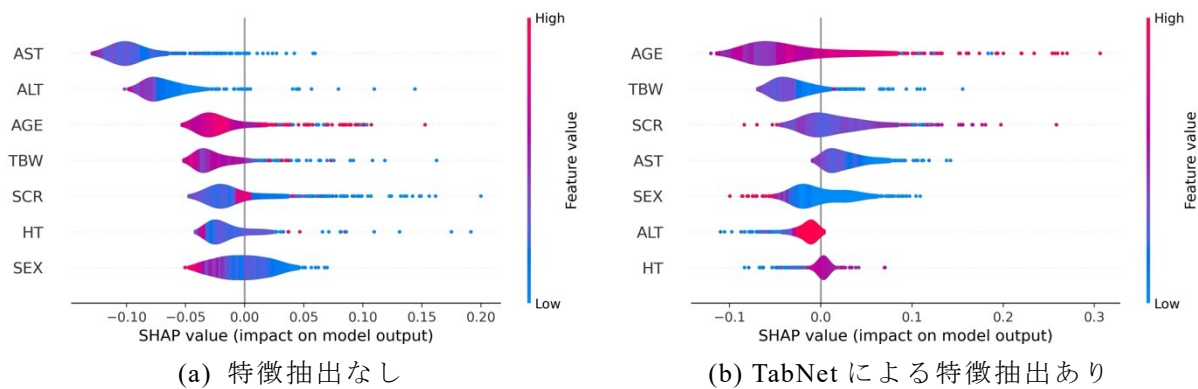


図 3. TVAE が生成したデータに対する異常検知の判断根拠外観

表 6 の CTGAN の結果に注目すると、特徴抽出を行うことで、僅かながら異常と判定されるデータの件数が増加した。それぞれのデータを比較した際に、特徴抽出前後で重複して異常とされたデータは 1 件のみであった。図 4 に、特徴抽出なしでは異常とされなかったが、TabNet による特徴抽出後は異常と判定されたデータに対する判断根拠の可視化の例を示す。横軸は SHAP 値を表しており、SHAP 値の性質として、各属性の SHAP 値と SHAP 値のベース値（平均値）を加えた値が、異常検知 AI の出力（予測値）と一致する。今回示したデータは、現実にはあり得ない体重-4.2 kg というデータであり、明らかに異常である。特徴抽出なしの状態でも、体重の異常さを SHAP 値に表せているが、他の属性の異常度が低く、最終的には異常度 38%と出力されていた。一方、特徴抽出を行うことで、異常度 81.5%と予測できており、特徴抽出が、合成データの品質評価に貢献していることが確認できる。

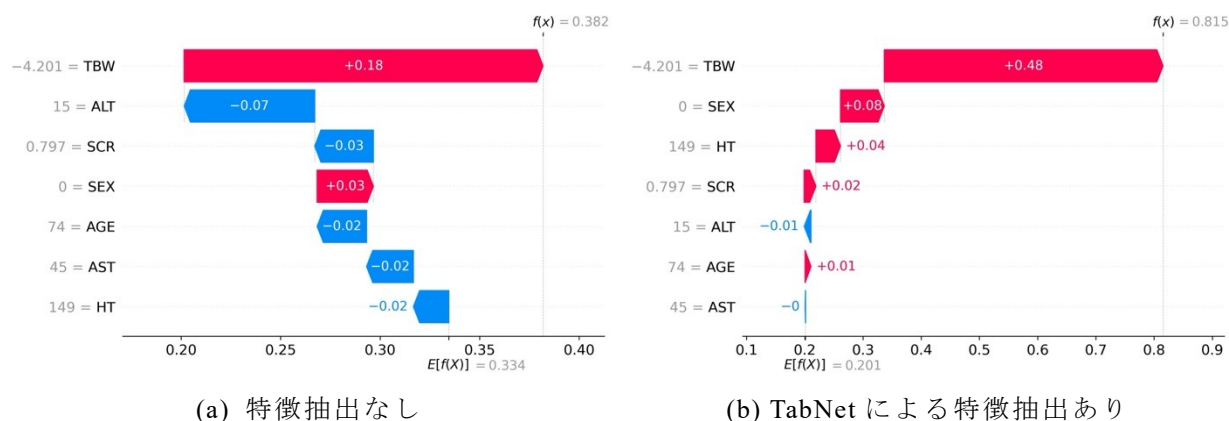


図 4. CTGAN が生成したデータに対する異常検知の判断根拠可視化の例 1

表 6 の TabDDPM の結果に注目すると、特徴抽出により、異常と出力されるデータ件数が 30 件程度増加していることが確認できる。3 章の評価では、高品質なデータを生成できていたにもかかわらず、異常と判定されるデータが多くなった要因は、学習データに含まれる外れ値のような値も生成できる、再現能力の高さに起因している。事実、異常と判定されたデータの多くはいずれかの属性の最大値・最小値を含んでいた。

図 5 に、特徴抽出前は異常度 53.7%であったものが、特徴抽出により異常度 79.2%まで増加したデータに対する、判断根拠の可視化結果を示す。この例は、年齢 1 歳、体重 164.1 kg という

う、それぞれ属性の最大値をもったデータである。組み合わせを考慮した際に明らかに異常であるが、特徴抽出なしの状態では、他の属性が適正範囲であることが影響し、正しく評価できなかった。一方、TabNet による特徴抽出を行うことで、異常度 79.2%と判定できた。

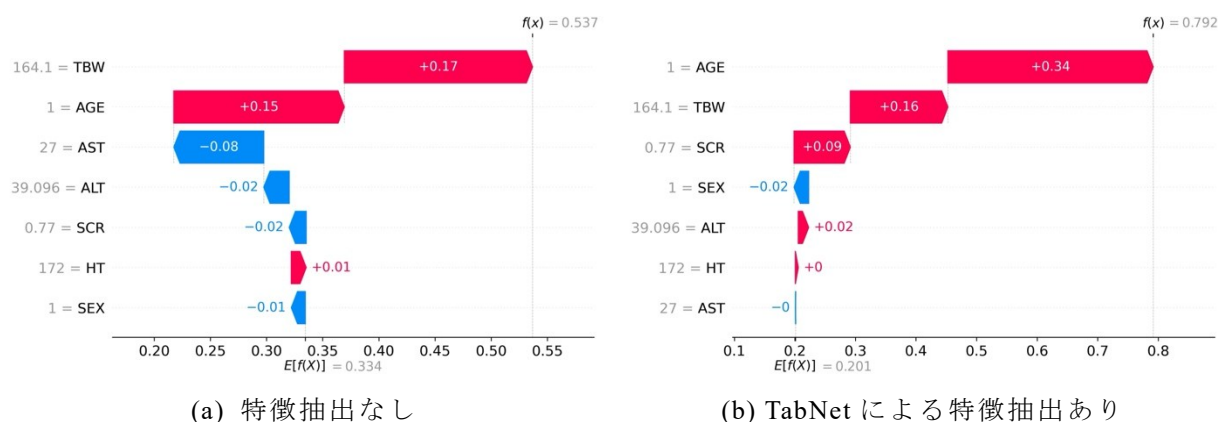


図 5. TabDDPM が生成したデータに対する異常検知の判断根拠可視化の例 1

図 6 に、特徴抽出前は異常度 76.7%であったものが、特徴抽出により異常度 49.9%まで減少したデータに対する、判断根拠の可視化結果を示す。この例は、体重 164.1 kg、身長 191 cm という、それぞれ属性の最大値をもったデータである。身長が高ければ、体重も増加する傾向にあるため、外れ値ではあるものの組み合わせを考慮した際に確実に異常とは言い切れないようなデータと考えられる。特徴抽出なしの状態では、体重と身長がそれぞれ異常度を高める方向に大きく寄与している。一方、TabNet による特徴抽出を行うことで、身長の SHAP 値が低下した。

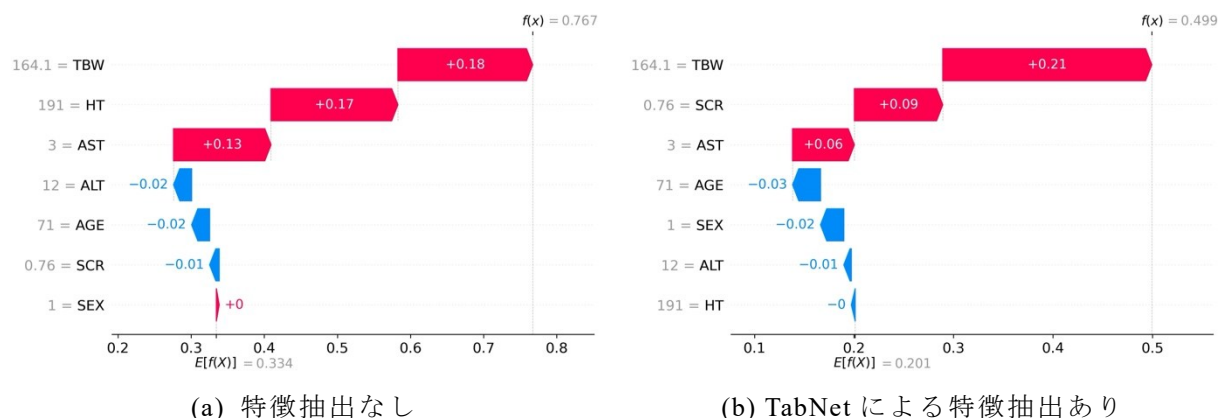


図 6. TabDDPM が生成したデータに対する異常検知の判断根拠可視化の例 2

以上の結果より、特徴抽出 AI と異常検知 AI を組み合わせた表形式医療データの品質評価を行う AI の有効性が示唆された。また、XAI の一種である SHAP により判断根拠の可視化を行うことで、本研究で提案した品質評価 AI の妥当性を示すことができた。

5. まとめ

本研究は、医薬品開発の効率化、臨床研究の高度化や、オープンな医療ビッグデータの実現を目的に、AIを用いた表形式医療データの合成と、合成データの品質評価に取り組んだ。特徴抽出AIと異常検知AI、そして説明可能なAIを組み合わせた品質評価手法を提案し、表形式データを合成する3種の生成モデルに対して検証を行なった。結果より、特徴抽出を行うことで、属性間の組み合わせを考慮することが可能となり、異常検知AIにより、本物らしさを評価できる可能性が示唆された。また、説明可能なAIによる異常検知AIの判断根拠可視化結果から、品質評価手法の妥当性を示した。

謝辞

本研究を遂行するにあたり、公益財団法人天野工業技術研究所から多大なご支援を頂きました。ここに記して謝意を示します。

参考文献

- 1) HOLFORD, N.; MA, S. C.; PLOEGER, B. A. Clinical trial simulation: A review. *Clinical Pharmacology & Therapeutics*, 2010, 88.2: 166-182.
- 2) SMANIA, Giovanni; JONSSON, E. Niclas. Conditional distribution modeling as an alternative method for covariates simulation: comparison with joint multivariate normal and bootstrap techniques. *CPT: pharmacometrics & systems pharmacology*, 2021, 10.4: 330-339.
- 3) XU, Lei, et al. Modeling Tabular Data Using Conditional GAN. *Advances in neural information processing systems*, 2019, 32.
- 4) KOTELNIKOV, Akim, et al. TabDDPM: Modelling Tabular Data with Diffusion Models. In: *International Conference on Machine Learning*. PMLR, 2023. p. 17564-17579.
- 5) ARIK, Sercan Ö and PFISTER, Tomas. TabNet: Attentive Interpretable Tabular Learning. In: *Proceedings of the AAAI Conference on Artificial intelligence*. 2021. p. 6679-6687.
- 6) LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. Isolation Forest. In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008. p. 413-422.
- 7) LUNDBERG, Scott M and SU-IN, Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of Advances in Neural Information Processing Systems 30*. NIPS, 2017.
- 8) OGAMI, Chika, et al. An artificial neural network– pharmacokinetic model and its interpretation using Shapley additive explanations. *CPT: pharmacometrics & systems pharmacology*, 2021, 10.7: 760-768.
- 9) DU, Yuntao and LI, Ninghu. Systematic Assessment of Tabular Data Synthesis Algorithms. *arXiv preprint arXiv:2402.06806*, 2024.
- 10) CHEN, Sihan, et al. PyOD 2: A Python Library for Outlier Detection with LLM-powered Model Selection. *arXiv preprint arXiv:2412.12154*, 2024.
- 11) JOSEPH, Manu. Pytorch Tabular: A Framework for Deep Learning with Tabular Data. *arXiv preprint arXiv:2104.13638*, 2021.