補完精度と適用可能割合を考慮した電子カルテデータの欠測値補完法の検討

多田 亜彩美† Berjab Nesrine†† Le Hieu Hanh†

† お茶の水女子大学共創工学部文化情報工学科 〒 112–8610 東京都文京区大塚 2-1-1
†† 東京科学大学情報理工学院 〒 152–8550 東京都目黒区大岡山 2-12-1
E-mail: † k2500004@edu.cc.ocha.ac.jp, le@is.ocha.ac.jp, ††berjab@c.titech.ac.jp

あらまし 近年，電子カルテデータを解析し医療支援等に用いる二次利用が注目されている．電子カルテデータには欠測が含まれる場合があるが，安易な欠測データの除去は解析結果の信頼性に影響を与えることがあるため，統計的な単変量の補完法が用いられているほか，電子カルテデータに対する補完手法も盛んに提案されている．従来の手法では数値に対する補完精度に注目することが多いが，実用上はすべての欠測箇所への対応が必要になり，全欠測数に対し手法をどの程度の割合で適用できるかという適用可能割合を含めた実用性の観点での検討が十分ではない．また電子カルテデータにおいては，医療上の判断に影響する検査結果の異常性に対する補完精度が重要になる．そこで本研究では，先行評価により高い補完精度が確認された線形補間法に注目し，他の手法を組合せ，異常性に対する補完精度及び適用可能割合という実用性の観点から補完手法の改善を検討した．具体的には，欠測を含む実際の電子カルテデータから血液検査結果を用いて，手法の補完精度と適用可能割合を求め，電子カルテデータに対する補完手法の実用性を評価した．

キーワード 医療，電子カルテデータ，欠測値

1 はじめに

1.1 背景

現在，厚生労働省による医療 DX の推進 [11] などを背景として，電子カルテデータを解析し医療支援等に用いる，電子カルテデータの二次利用が注目されている．一方で，電子カルテデータには欠測が含まれる場合がある．後述する，米国で収集された実際の病院患者のデータベース MIMIC-IV [1], [3], [4] において，検体検査結果データを調べると全体で約 10.14%の欠測があった．

電子カルテデータの分析例として，医療指示列の典型例であるクリニカルパスにおける分岐について，年齢や体重などの患者情報を説明変数としたロジスティック回帰分析から，分岐要因を推定した研究がある [14]．この研究では，多様な分岐要因に対応するため，用いる患者情報の種類を増やす必要性が示唆された．検体検査結果は医療上の判断時に患者の状態を確認するための重要な情報 [12] であるから，用いる患者情報として検体検査結果が候補に挙げられる．しかし先述の通り，検体検査結果をはじめとして電子カルテデータには欠測が含まれる場合があり，分析に用いるためには欠測への対応が必要である．安易な欠測の除去はデータの偏りに繋がり，解析結果の信頼性に影響を与える場合がある．そのため，欠測箇所に適切な値を代入し補完する手法が必要になる．

補完法は，いくつかの観点で分類することができる．Sun ら [6] は，伝統的な統計的手法，機械学習手法，深層学習手法の 3 つに補完法を分類している．また別の観点として，Jazayeri ら [2] が言及するように，欠測箇所を持つ変量と同じ変量のみを考慮して補完する単変量の場合と，欠測箇所を持つ変量に加え他の変量も考慮して補完する多変量の 2 つの場合に分類することもできる．

電子カルテデータにおいて，一般の入院患者や外来患者では欠測率が高くなると考えられるが [5]，このような疎なデータセットを用いた機械学習では補完精度を低下させる可能性がある [7]．そこで本研究では，統計的な補完法に着目している．

応用研究や実務の場面で用いられることが多いとされる平均値代入法 [10] を例に挙げると，この手法は統計的な単変量の補完法に分類できる．Sun ら [6] は，よく用いられる補完法の評価を整理しており，平均値代入法については，単純で実装が簡単であり，処理が速いことを利点として挙げている．平均値代入法に限らず，この特徴は統計的な単変量の補完法に一般に当てはまると考えられ，このような実用上の利用のしやすさという点からも，筆者らは統計的な単変量の補完法に着目している．

補完法の評価の視点として，数値に対する補完精度に着目されることが多いが，電子カルテデータにおいては，数値そのものではなく，医療上の判断に影響する数値の異常性が重要になる場合がある．また，実用上はすべての欠測箇所への対応が必要になるが，全欠測数に対し手法をどの程度の割合で適用できるかという適用可能割合を含めた実用性の観点での検討は十分ではない．

そこで筆者らは，統計的な単変量の補完法である前方補完法，後方補完法，平均値代入法，線形補間法について，異常性に対する補完精度と，全欠測数に対して補完法を適用できた割合を適用可能割合と定義し評価を行った [9]．その結果，線形補間法

について，補完精度が比較的高い一方，適用可能割合が比較的低いという課題が明らかになった．

1.2 目　　的

以上の背景を踏まえ，本研究では，筆者らの先行評価 [9] により，異常性に対する補完精度が比較的高い一方で適用可能割合に課題があった線形補間法に着目し，前方補完法や後方補完法などと組合せることで，適用可能割合の改善と，補完精度のさらなる向上を検討した．

具体的には，実際の電子カルテデータである MIMIC-IV の血液検査結果へ補完法を適用し，全欠測数に対して補完法を適用できた割合である適用可能割合と，補完値から判断した異常性に対する補完精度を評価した．

なお，本研究で用いたデータセットの特徴として，補完対象となる欠測箇所が概ね全て異常と判断されるデータであったため，そのデータの中で補完値から異常と予測できた割合である再現率を，本研究では異常性に対する補完精度として重視した．

1.3 本研究の貢献

本研究の主な貢献は，先行評価により適用可能割合に課題があった線形補間法に着目し，線形補間法が適用できない欠測箇所を隣接する値で補完する前方補完法または後方補完法を組合せて適用し，実用上重要となる補完精度と適用可能割合の観点から，実際の電子カルテデータを用いて評価している点である．

結果は，線形補間法を単独で使用した場合に比べて，前方補完法と後方補完法の両方を組合せた場合，適用可能割合は約 24.15%上昇し，改善が確認された．一方，異常性は欠測の前後で大きく変わらない場合が多いことが推測され，補完精度として重視した再現率については，手法の組合せによる大きな変化はないことが確認された．

1.4 本稿の構成

本稿は，まず第 2 節で関連研究について述べる．第 3 節で本研究で検討した手法について説明し，次に第 4 節で実験に用いたデータセットと評価方法，そして実験結果について述べる．最後に第 5 節にてまとめと今後の課題を述べる．

2 関 連 研 究

2.1 統計的な単変量の補完法

従来から用いられている，統計的な単変量の補完法としては次のようなものがある．

LOCF（Last Observation Carried Forward）法は，ある変量において，観測された最後の値，言い換えれば，欠測する直前の値を欠測箇所に代入する方法である．医薬品開発の臨床試験において，試験の途中で対象者が脱落し，欠測が生じた場合などによく用いられてきた [10]．

LOCF 法の逆の発想として，欠測の直後の値を欠測箇所に代入する方法もある．

1.1 にて例として挙げた平均値代入法は，ある変量において，他の観測値から計算された平均値を欠測箇所に代入する方法である．平均値以外に，中央値や最頻値などの統計量を用いる場合もある．

線形補間法は，ある変量において，欠測箇所の前後 2 点を通る直線を求め，欠測箇所の値は，その前後 2 点が作る線分上にあるものとして，直線の式から求めた値を欠測箇所に代入する手法である．

なおここでの「補間」という語は，2 点の間の値を補う内挿手法を表すものとして用いており,「補完」とは区別をしている．本稿では,「補間」は線形補間などの内挿手法を指し,「補完」は欠測箇所への値の代入処理全般を指すものとする．

2.2 電子カルテデータに対する補完法

電子カルテデータに対する補完法は，近年盛んに提案されている．

兵頭ら [13] は，臨床的な活用頻度が低くなっているという蛋白分画検査に着目した．蛋白分画検査の波形情報と，患者の年齢・性別の情報から，血液検査の結果を予測する機械学習モデルを構築し，その有効性を示した．なおこの研究では，年齢・性別に応じた基準範囲を用いて，血液検査の結果値を「基準範囲内」と「基準範囲外」の 2 値データに変換し，その識別性能を評価している．

Luo [5] は，電子カルテデータに対する欠測値補完の共有タスクチャレンジ the Data Analytics Challenge on Missing data Imputation (DACMI) に参加した 12 のチームが提案した補完法とその結果を報告している．このチャレンジでは，実際の電子カルテデータから抽出された 13 種の血液検査結果と，検査が行われたタイムポイントから成るデータセットに対し，12 のチームが様々な方法で値の補完を試みた．人工的に発生させた欠測を用いて，nRMSD により結果を評価している．

このチャレンジの中で総合的に最もよい nRMSD を達成したのが，Xu ら [7] が提案した手法であった．時間情報と，変数である 13 種の血液検査を横断する情報から特徴量を構築し，機械学習による補完法を提案した．

統計的な手法として，Jazayeri ら [2] は，患者間の類似性を考慮した補完手法を提案した．13 種の血液検査結果から患者間のユークリッド距離を計算し，時間間隔を考慮した上で患者間の類似度を求め，それを重みとして他の患者の結果値を用いた加重平均により補完した．

DACMI では欠測率が比較的低い ICU 入院患者のデータセットが用いられているが，Luo [5] が課題として指摘している通り，一般の入院患者や外来患者では欠測率が高くなると考えられる．このような欠測率の高い変数が特徴量に含まれる場合，機械学習での補完精度を低下させる可能性があることを Xu ら [7] が課題として指摘している．そこで本研究では，統計的なアプローチで欠測値補完を行うこととした．

Luo [5] が報告したチャレンジでは評価に nRMSD が用いられていた通り，一般に，補完法の評価は数値に対する補完精度に着目することが多い．一方で，兵頭ら [13] のように，結果値を基準範囲内と基準範囲外の 2 値データに変換し，その識別性能を評価するという視点もある．電子カルテデータにおいては，

データ例

行番号	患者ID	時間	値
1	0001	2010-01-07 10:03:00	欠測
2	0001	2010-01-09 11:20:00	0.20
3	0001	2010-01-10 13:55:00	0.30
4	0001	2010-01-11 15:40:00	欠測
5	0001	2010-01-15 20:23:00	0.50
6	0001	2010-02-10 13:37:00	0.60
7	0001	2010-03-15 14:05:00	欠測

値を補完 ➡

前方補完法（F法）	後方補完法（B法）	線形補間法（LI法）	時間間隔を考慮した線形補間法（t-LI法）
適用不可	0.20	適用不可	適用不可
0.20	0.20	0.20	0.20
0.30	0.30	0.30	0.30
0.30	0.50	0.40	0.34
0.50	0.50	0.50	0.50
0.60	0.60	0.60	0.60
0.60	適用不可	適用不可	適用不可

図 1　各手法の適用例

数値そのものではなく，医療上の判断に影響する数値の異常性が重要になる場合があることから，本研究では，結果の異常性に対する補完精度を評価の指標として用いることとした．

また，Jazayeri ら [2] の研究では，十分な類似患者がデータセットの中にないと判断された場合は，線形補間法を併用している．このように，補完法によっては手法の適用ができない欠測箇所が存在することもある．実用上はすべての欠測箇所への対応が必要になるが，全欠測数に対し手法をどの程度の割合で適用できるかという適用可能割合の評価は十分にされていないと考え，本研究では，適用可能割合も評価値として用いた．

3　検 討 手 法

本研究では，線形補間法に着目し，線形補間法が適用できない場合に他の手法を組合せることで，適用可能割合の改善と，補完精度のさらなる向上を検討した．

本節では，検討した手法の詳細を述べる．基本となる 4 種類の手法について 3.1〜3.4 にて説明し，3.5 にて手法の組合せについて説明する．

また，次の 3.1〜3.4 にて説明する 4 種類の手法については，架空のデータ例を用いて，図 1 に各手法の適用例を示した．

3.1　前方補完法

患者ごとにグループ化したデータに対し，欠測箇所の直前の値を欠測箇所に代入する．2 つ以上連続して欠測箇所がある場合は，連続して同じ値を代入する．適用例は図 1 に示す通りである．

一般には LOCF 法と呼称される手法であるが，本稿では，この手法を「前方補完法」と呼ぶ．また略称として，本稿では「F 法」と表現する．

3.2　後方補完法

患者ごとにグループ化したデータに対し，欠測箇所の直後の値を欠測箇所に代入する．2 つ以上連続して欠測箇所がある場合は，連続して同じ値を代入する．適用例は図 1 に示す通りである．

本稿では，この手法を「後方補完法」と呼ぶ．また略称として，本稿では「B 法」と表現する．

3.3　線形補間法

患者ごとにグループ化したデータに対し，欠測箇所の前 (x_1, y_1) と後 (x_2, y_2) の 2 点を通る直線 (1) を考え，欠測箇所の点 (x_m, y_m) はその前後 2 点が作る線分上にあるものとして求め，値 y_m を欠測箇所に代入する．この時，欠測箇所とその前後 2 点は線分上に等間隔に並んでいるものとする．

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) \tag{1}$$

また，本研究では，欠測箇所の前後 2 点が揃い，内挿を行える場合のみ線形補間法を適用する．適用例は図 1 に示す通りである．

略称として，本稿では「LI 法」と表現する．

3.4　時間間隔を考慮した線形補間法

3.3 と同様に，患者ごとにグループ化したデータに対し，欠測箇所の前 (x_1, y_1) と後 (x_2, y_2) の 2 点を通る直線 (1) を考え，欠測箇所の点 (x_m, y_m) はその前後 2 点が作る線分上にあるものとして求め，値 y_m を欠測箇所に代入する．この時，欠測箇所の前後の 2 点と欠測箇所の点の x 座標は，それぞれが持つ時間情報 $time_1$，$time_2$，$time_m$ とし，時間間隔を考慮して欠測箇所の値 y_m を求める．

また，本研究では，欠測箇所の前後 2 点が揃い，内挿を行える場合のみこの手法を適用する．適用例は図 1 に示す通りである．

略称として，本稿では「t-LI 法」と表現する．

3.5　手法の組合せ

本研究では，内挿を行える場合にのみ線形補間法を適用しており，図 1 の適用例に示した通り，欠測が一連のデータの端点にある場合などには，手法を適用することができない．

そこで，線形補間法が適用できない欠測箇所を隣接する値で補完するという実用性を意識した方針として，前方補完法または後方補完法を組合せた．可能な限り線形補間法を適用し，欠測箇所の直前の点が存在しない場合は後方補完法を，欠測箇所の直後の点が存在しない場合は前方補完法を組合せて適用することで，手法の適用可能割合の改善を試みた．

手法の組合せ方の一覧は以下の通り．

- 線形補間法＋前方補完法 (LI+F 法)

表 1　各手法の適用可否

欠測の状況	データ例			手法の適用可否									
	行番号	患者 ID	値	F 法	B 法	LI 法	t-LI 法	LI 法			t-LI 法		
								+F 法	+B 法	+F 法+B 法	+F 法	+B 法	+F 法+B 法
すべての値が欠測	1	0001	欠測	×	×	×	×	×	×	×	×	×	×
	2	0001	欠測										
欠測箇所の前に値なし	3	0002	欠測	×	○	×	×	×	○	○	×	○	○
	4	0002	3.6										
	5	0002	3.3										
	6	0002	4.1										
	7	0002	4.3										
欠測箇所の後に値なし	8	0003	3.0	○	×	×	×	○	×	○	○	×	○
	9	0003	3.2										
	10	0003	欠測										
欠測箇所の前後に値あり	11	0004	3.1	○	○	○	○	○	○	○	○	○	○
	12	0004	欠測										
	13	0004	3.0										
	14	0004	3.8										

- 線形補間法＋後方補完法 (LI+B 法)
- 線形補間法＋前方補完法＋後方補完法 (LI+F+B 法)
- 時間間隔を考慮した線形補間法＋前方補完法 (t-LI+F 法)
- 時間間隔を考慮した線形補間法＋後方補完法 (t-LI+B 法)
- 時間間隔を考慮した線形補間法＋前方補完法＋後方補完法 (t-LI+F+B 法)

架空のデータ例を用いて，欠測の状況による手法の適用可否を表 1 に整理した．この表は，欠測の状況と各手法の適用可否の対応関係を整理したものである．例えば，行番号 3〜7 の患者 ID が 0002 の患者が持つ欠測個所については，欠測箇所が最初の行にあり，前の値がないため，線形補間法を適用することができない．そのような場合に，欠測箇所の直後の値を欠測箇所に代入する後方補完法を組合せることで，線形補間法のみでは補完することができなかった欠測個所に対応が可能となる．

なお，表 1 の例では，簡単のために，欠測の状況と患者ごとにグループ化したデータを 1 対 1 に対応させているが，実際には，1 つのグループの中に複数の状況の欠測が発生している場合がある．

4　評 価 実 験

本節では，実験に使用したデータセット，評価方法について説明し，最後に実験結果を述べる．

4.1　データセット

実験には，米国で収集された，実際の病院患者のデータベース MIMIC-IV から，検体検査結果を記録した labevents テーブルから抽出したデータを用いた．実験に用いたデータセットの概要は表 2 のとおりである [8].

検査結果値が記録された valuenum 列に欠測が含まれる場合があり，本実験で補完を行った．また，flag 列には，検査結果が基準範囲を超えるなど異常を示す場合には，abnormal の文字列が記録されている．それ以外の場合は何も記録されず，NULL となっている．本実験では結果の異常性に着目しているため，この flag 列を正解セットとして用いた．詳細は次節で述べる．なお，labevents テーブルには，主に入院患者の検体検査結果が記録されているが，外来患者のデータも含まれている．そのため，storetime に記録されている時間の間隔は一定ではない．

表 2　実験に用いたデータセットの各列の概要

列名	説明
labevent_id	テーブル内の各行にユニークな id
subject_id	患者ごとにユニークな id
itemid	検査項目ごとにユニークな id
storetime	検査結果が利用可能になった時間
valuenum	検査結果を示す数値データ
valueuom	検査結果値に対する単位
ref_range_lower	検査結果の基準値における下限
ref_range_upper	検査結果の基準値における上限
flag	検査結果の異常を示す文字列
comments	匿名化された自由記述のコメント．結果やその解釈に関する注意や，場合によっては検査結果そのものがコメントに含まれる場合がある．

4.1.1　正解セットの構築と前処理

valuenum 列に記録された検査結果値が欠測しており，基準値と比較する数値がない場合にも，flag 列に異常を示す abnormal の文字列が記録されていることがあった．このような行の comments 列の内容を抽出して調べると，数値としての検査結果が得られず valuenum 列は欠測しているが，何らかの異常を示す兆候から，flag 列に abnormal が記録される場合があることが確認された [8]．このことから，valuenum 列の値を補完した後，基準値と補完値を比較し判断した結果の異常性を，flag 列と比較することで評価できると考え，flag 列を正解セットとして用いることとした．

表 3　データセットの概要

検査項目名	全データ数	欠測数	欠測率	欠測箇所における異常数	欠測箇所における異常率
アルブミン	1,032,387	51	0.005%	51	100.00%
血小板	4,199,386	1,499	0.036%	1,499	100.00%
クレアチニン	4,334,840	938	0.022%	938	100.00%
尿素窒素	4,203,472	1,544	0.037%	1,544	100.00%
尿酸	192,219	98	0.051%	98	100.00%
アルカリホスファターゼ	1,602,488	33	0.002%	33	100.00%
カリウム	4,481,959	384	0.009%	384	100.00%
C 反応性蛋白	177,842	3,573	2.009%	3,368	94.26%
グルコース	3,903,505	860	0.022%	860	100.00%
クレアチンホスホキナーゼ	334,663	72	0.022%	72	100.00%

正解セット構築に関連した前処理として，次の 2 つの処理を実施した．

まず，comments 列に記録された内容から，結果が異常を示したことが推測されるが，flag 列には何も記録されていない場合もあったため，その場合は flag 列に abnormal の文字列を補った．

次に，上記のように comments 列の内容から異常性を判断することができないため，valuenum 列が欠測している行のうち，comments 列の内容が NULL であるか，あるいは完全に匿名化されている場合は，データセットから除外した．

4.1.2　使用した検査項目

labevents テーブルには様々な検体検査結果が記録されているが，本実験ではまず，兵頭ら [13] が用いた検査実施の頻度が高い 21 項目の血液検査結果を抽出した．

これらの 21 項目に対し 4.1.1 に述べた前処理を行った結果，valuenum 列が欠測している行がすべて除外されてしまうなど，実験に用いることができない検査項目が 11 項目あった．本実験ではその 11 項目を除き，10 の検査項目を実験に用いることとした．データセットの概要は表 3 の通りである [9]．

4.1.3　データセットの特徴と妥当性

本実験に用いたデータセットの特徴として，4.1.1 に述べた前処理の結果，概ね全ての欠測箇所の flag 列には abnormal の文字列が記録され，表 3 に示す通り，欠測箇所は概ね全て異常と判断されるデータとなった．本研究の目的は，異常性に対する補完精度の評価であり，これらの特徴はその目的に適しているため，このデータセットを用いて評価を行うこととした．

4.2　検討手法の一覧と実装

検討した手法の一覧は以下の通り．

1. 前方補完法 (F 法)
2. 後方補完法 (B 法)
3. 線形補間法 (LI 法)
4. 線形補間法＋前方補完法 (LI+F 法)
5. 線形補間法＋後方補完法 (LI+B 法)
6. 線形補間法＋前方補完法＋後方補完法 (LI+F+B 法)
7. 時間間隔を考慮した線形補間法 (t-LI 法)
8. 時間間隔を考慮した線形補間法＋前方補完法 (t-LI+F 法)
9. 時間間隔を考慮した線形補間法＋後方補完法 (t-LI+B 法)
10. 時間間隔を考慮した線形補間法＋前方補完法＋後方補完法 (t-LI+F+B 法)

表 4　使用したメソッドとパラメータ設定

手法	メソッド	パラメータ
1.F 法	ffill	default
2.B 法	bfill	default
3.LI 法	interpolate	limit_area="inside"
4.LI+F 法	interpolate	limit_direction="forward"
5.LI+B 法	interpolate	limit_direction="backward"
6.LI+F+B 法	interpolate	limit_direction="both"
7.t-LI 法	interpolate	method="time", limit_area="inside"
8.t-LI+F 法	interpolate	method="time", limit_direction="forward"
9.t-LI+B 法	interpolate	method="time", limit_direction="backward"
10.t-LI+F+B 法	interpolate	method="time", limit_direction="both"

検討手法の実装には pandas を用いた．使用したメソッドとパラメータ設定は表 4 の通りである．明記のないパラメータについては，デフォルトの設定で使用した．

なお，線形補間を行う interpolate メソッドでは，内挿が行えない箇所は隣接する値が代入されるため，補完方向を指定するパラメータ limit_direction を表 4 の通りに設定することで，内挿が行えない欠測箇所に組合せる補完方法を指定した．

4.3　評 価 方 法

評価指標として次の 2 つを用いた．

4.3.1　適用可能割合

表 1 に整理した通り，欠測の状況により手法の適用可否が異なる．検査項目ごとに，全欠測数に対して補完法を適用できた割合を，適用可能割合として求めた．

4.3.2　補 完 精 度

電子カルテデータでは，医療上の判断に影響する結果の異常性が重要になる場合があることから，本研究では，補完値から判断した異常性に対する適合率，再現率を求めた．

4.1.1 に示した通り，flag 列を検査時に異常と判断された結果を示す正解セットとして用いる．各手法を適用し補完値を求めた後，基準範囲と比較し，基準範囲外であれば，新たに作成

した列に異常を示す abnormal を記録した．この結果を 4.1.1 で構築した正解セットと比較し，適合率，再現率を求めた．本実験の場合は，適合率は，補完値から異常と予測した中で，実際に異常であった割合であり，再現率は，実際に異常であった中で，補完値から異常と予測できた割合である．

適合率について，4.1.3 で述べた通り，本実験で用いたデータの欠測箇所は，概ね全て異常と判断されるデータであったため，補完値から異常であると予測した場合，必ず正解となる．そこで本実験では，異常性に対する補完精度として，再現率を重視して評価を行った．

4.4 実験環境

本実験は，Python（バージョン 3.12.3）で実行し，検討手法の実装などデータ処理のためのライブラリとして pandas（バージョン 2.2.3）を用いた．

4.5 実験結果

10 の検査項目に対し各検討手法を適用した結果，適用可能割合は図 2 に示す通りとなった．補完精度については，適合率は図 3 に，再現率は図 4 に示す通りとなった．検査項目ごとに適用可能割合，適合率，再現率を算出し，その分布を図 2，3，4 にまとめている．

適用可能割合については，6. LI+F+B 法，10. t-LI+F+B 法のように，線形補間法の適用ができない場合に前方補完法と後方補完法の両方を組合せた方法が，最も適用可能割合が高かった．3. LI 法または 7. t-LI 法を単独で使用した場合に対して，6. LI+F+B 法，10. t-LI+F+B 法はいずれも約 24.15% 適用可能割合が改善された．

次に，図 3 に示した通り，適合率は全体で平均約 1.00 となった．4.3 で述べた通り，欠測箇所は概ね全て異常と判断されるデータであり，補完値から異常であると予測した場合，必ず正解となるため，このような結果となった．

そこで，ここからは再現率を用いて補完精度について考察を行う．再現率は，本実験では，実際に異常と判断されたデータの中で，補完値から異常と予測できた割合である．

再現率は，全体で平均約 0.78 となった．手法の組合せによる，補完精度の大きな変化はなかった．実際のデータを確認すると，異常性は欠測の前後で大きく変わらない場合が多いことが推測され，線形補間法，前方補完法，後方補完法は，いずれも欠測箇所に隣接した値を用いて補完しているため，このような結果になったと考えられる．

また，3〜6 の線形補間法と，7〜10 の時間間隔を考慮した線形補間法の結果をそれぞれ比較すると，時間間隔を考慮した場合の方が平均約 3.12%再現率が上昇した．本実験で使用したデータセットには，入院時だけでなく外来時の検査データも含まれ，検査の時間間隔が一定ではないため，時間間隔を考慮することにより補完精度が上昇したと考えられる [9]．

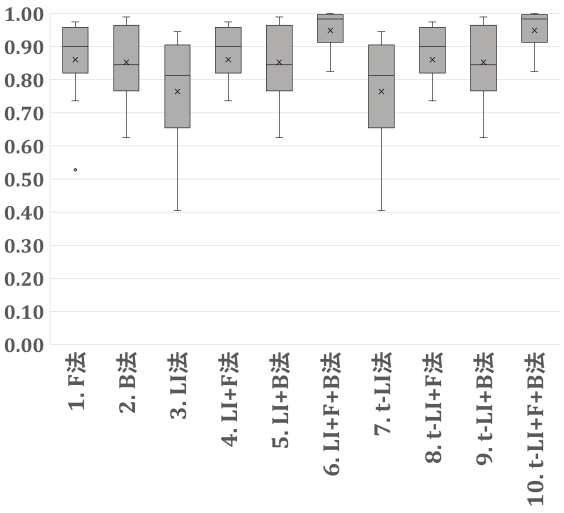

図 2　適用可能割合の結果

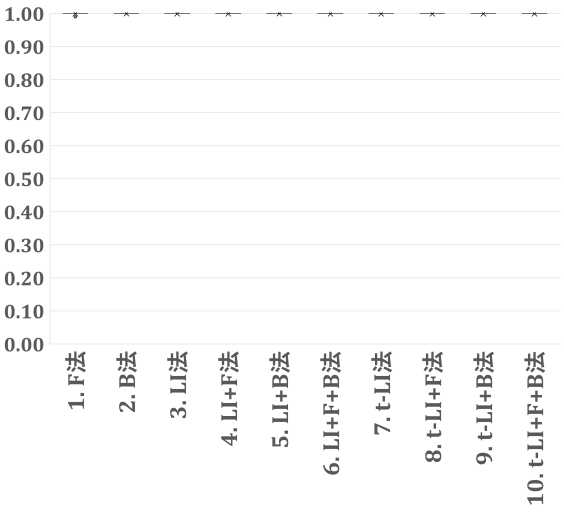

図 3　補完精度（適合率）の結果

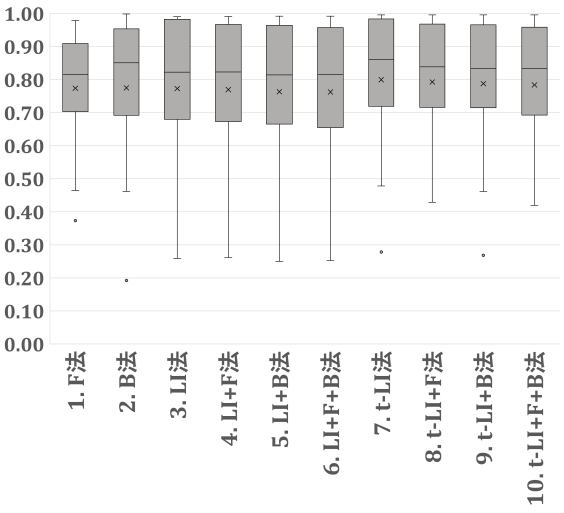

図 4　補完精度（再現率）の結果

5 おわりに

5.1 まとめ

本研究では，数値の異常性に対する補完精度と，手法の適用可能割合という点から，統計的な単変量の補完法について検討を行った．特に，線形補間法に着目し，前方補完法や後方補完法と組合せることで，適用可能割合の改善と，補完精度のさらなる向上を検討した．結果は，線形補間法を単独で使用した場合に比べて，前方補完法と後方補完法の両方を組合せた場合，適用可能割合は約 24.15%上昇し，改善が確認された．実際の運用ではすべての欠測に対応する必要があり，適用可能割合の改善は実用上重要である．一方で，異常性は欠測の前後で大きく変わらない場合が多く，補完精度として重視した再現率については，手法の組合せによる大きな変化はなかった．

5.2 今後の課題

今後も線形補間法に着目し，適用可能割合の改善と，補完精度のさらなる向上を検討する．本研究では，実用的な利用のしやすさという点で単変量の補完法に限定して検討を行ったが，補完精度の向上を目指し，検査項目間の相関関係や患者間の類似性を考慮した手法の組合せを検討する．さらに，本研究の評価では結果の異常性に着目したが，今後，補完値そのものに対する精度の評価も行う．

文　献

[1] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, R. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H.E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, Vol. 101, No. 23, pp. e215–e220, 2000.

[2] Ali Jazayeri, Ou Stella Liang, and Christopher C. Yang. Imputation of missing data in electronic health records based on patients' similarities. *Journal of Healthcare Informatics Research*, Vol. 4, pp. 295–307, 2020.

[3] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark. MIMIC-IV (version 3.1). PhysioNet, 2024.

[4] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, L. H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, Vol. 10, No. 1, 2023.

[5] Yuan Luo. Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, Vol. 23, No. 1, pp. 1–9, 2022.

[6] Yige Sun, Jing Li, Yifan Xu, Tingting Zhang, and Xiaofeng Wang. Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, Vol. 227, p. 120201, 2023.

[7] Xiao Xu, Xiaoshuang Liu, Yanni Kang, Xian Xu, Junmei Wang, Yuyao Sun, Quanhe Chen, Xiaoyu Jia, Xinyue Ma, Xiaoyan Meng, Xiang Li, and Guotong Xie. A multidirectional approach for missing value estimation in multivariate time series clinical data. *Journal of Healthcare Informatics Research*, Vol. 4, pp. 365–382, 2020.

[8] 多田亜彩美, Le Hieu Hanh. 電子カルテデータ解析における欠測値補完法の一検討. WebDB 夏のワークショップ 2025, No. 4C-1, 2025.

[9] 多田亜彩美, Berjab Nesrine, Le Hieu Hanh. 電子カルテデータ解析における欠測値補完法の実用性の評価. 情報処理学会第 88 回全国大会, 2026.

[10] 高井啓二, 星野崇宏, 野間久史. 欠測データの統計科学—医学と社会科学への応用. 調査観察データ解析の実際 1. 岩波書店, 2016.

[11] 厚生労働省. 医療 DX について. `https://www.mhlw.go.jp/stf/iryoudx.html`. 2026.2.11 アクセス.

[12] 横田治夫. 電子カルテデータ解析 医療支援のためのエビデンス・ベースド・アプローチ. 共立出版, 2022.

[13] 兵頭勇己, 久原太助, 檜山麻里子, 安井繁宏, 畠山豊, 奥原義保. 蛋白分画検査の波形情報を使用した血液検査の欠測値補間の試み. 日本医療情報学会 第 41 回医療情報学連合大会（第 22 回日本医療情報学会学術大会）, No. 3-G-1-01, 2021.

[14] 本田祐一, 山田達大, 萱原正彬, Le Hieu Hanh, 串間宗夫, 小川泰右, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫. 患者の固有情報及び動的状況を考慮したクリニカルパス分岐要因推定. In *DEIM Forum 2019*, No. D1-5, 2019.