6N-06

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

多田 亜彩美† Berjab Nesrine‡ Le Hieu Hanh†
お茶の水女子大学† 東京科学大学‡

1. はじめに

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期待されている．電子カルテデータには欠測が含まれる場合があり，安易な欠測の除去がデータの偏りにつながることがあるため LOCF（Last Observation Carried Forward）法や平均値代入法などの統計的な単変量代入法で値を補完する手法が用いられる．従来の手法は主に数値に関する補完精度の評価が中心であるが，電子カルテデータにおいては医療上の判断に関わる検査結果の異常性に着目した補完精度に加えて，全ての欠測箇所に対処が必要であるという適用可能割合の考慮が必要である．そこで本研究では，利用のしやすさという点から統計的な単変量代入法に焦点を当て，実際の電子カルテデータより欠測を含む血液検査結果値を用いて，結果の異常性に着目した補完精度と，適用可能割合という実用上の観点から評価した．本研究の主な貢献は，統計的な単変量代入法を実用性の観点から評価した点である．

2. 検討手法

本研究で検討した補完法は以下の通りである．先行評価の結果[1]から，補完精度の高かった平均値代入法，線形補間法に着目し，下記の手法を検討した．

(1) 前方補完法

欠測箇所より前の値のうち最も新しい値を欠測箇所に代入する．

(2) 後方補完法

欠測箇所より後の値のうち最も新しい値を欠測箇所に代入する．

(3) 平均値代入法

存在する値から平均値を求め欠測箇所に代入する．平均値の求め方は次の 4 種類を検討した．

(3)-1 平均値代入法（患者全体）

患者全体の値を用いて平均値を求めた．

(3)-2 平均値代入法（患者ごと）

患者ごとに全ての値を用いて平均値を求めた．

(3)-3 平均値代入法（欠測箇所前）

患者ごとに欠測箇所より前の値を用いて平均値を求めた．

(3)-4 平均値代入法（時間間隔を考慮）

患者ごとに時間間隔による重みから重み付き平均値を求めた．

(4) 線形補間法

欠測箇所の前後の値 2 点を通る直線を求め，欠測箇所の値は 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した．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適用した．また欠測箇所の前後の値および欠測箇所の線分上の位置として次の 2 種類を検討した．

(4)-1 線形補間法（等間隔）

線分上の位置は等間隔に設定した．

(4)-2 線形補間法（時間間隔を考慮）

線分上の位置は，欠測箇所の前後の値および欠測箇所の時間間隔の比に基づき設定した．

3. 評価実験

3.1. データセット

本研究では，米国で収集された実際の病院患者のデータベース MIMIC-IV[2],[3],[4]の，検体検査結果を記録した labevents テーブルより，表 1 に示す 10 の血液検査項目から成るデータセットを作成した．また，検査時点で異常と判断された場合に「abnormal」の文字列が記録される flag 列を正解セットとして用いることとした．なお，正解セット構築のための前処理として，異常を示すコメントが記録されている場合がある comments 列を参照し，flag 列への「abnormal」の追記と，異常の有無が明確に確認できないデータは除外する処理を実施した．

表 1 の通り，欠測箇所はほぼ全て異常と判断されるデータとなったが，本研究では異常性に対する補完精度に着目しているため，このデータセットで評価を行うこととした．

表１．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%
血小板	4,199,386	1,499	0.036%	1,499	100%
クレアチニン	4,334,840	938	0.022%	938	100%
尿素窒素	4,203,472	1,544	0.037%	1,544	100%
尿酸	192,219	98	0.051%	98	100%
アルカリホスファターゼ	1,602,488	33	0.002%	33	100%
カリウム	4,481,959	384	0.009%	384	100%
C反応性蛋白	177,842	3,573	2.009%	3,368	94%
グルコース	3,903,505	860	0.022%	860	100%
クレアチンホスホキナーゼ	334,663	72	0.022%	72	100%

Evaluating the practicality of missing data imputation methods for electronic medical records
†Asami Tada
‡Nesrine Berjab
†Hieu Hanh Le
†Ochanomizu University
‡Institute of Science Tokyo

Copyright ©2026 Information Processing Society of Japan.
All Rights Reserved.

3.2. 評価方法

次の 2 つの観点から各手法を評価した.

3.2.1. 適用可能割合

欠測の状況により手法の適用可否が異なるため，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.

3.2.2. 補完精度

電子カルテデータにおいては医療上の判断に関わる検査結果の異常性が重要になるため，本研究では補完精度を示す評価値として，補完値から判断した異常性に対する再現率を用いた. 各手法を適用後， MIMIC-IV 上で定義された当該検査項目の基準範囲から異常性を判定し，正解セットの flag 列と比較して再現率を求めた.

3.3. 実験結果

適用可能割合の結果は図 1 に，補完精度の結果は図 2 に示す通りである.

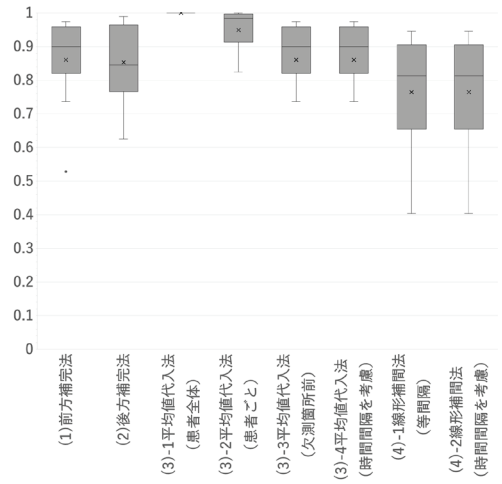

図 1. 適用可能割合の結果

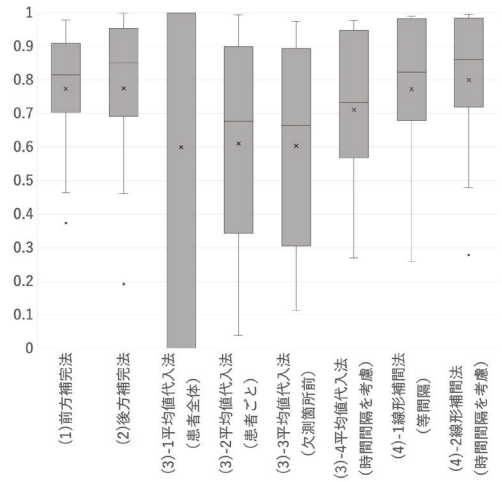

図 2. 補完精度（再現率）の結果

適用可能割合について，(3)-1 平均値代入法（患者全体）は全ての検査項目で 100%であり，(4)線形補間法は平均約 76%で比較的低い結果となった. (3)-1 平均値代入法（患者全体）はどのような欠測にも適用できる一方で，その他の手法は，欠測の前方あるいは後方に値がないと適用できないことから，このような結果となった.

補完精度については，(1)前方補完法，(2)後方補完法，(4)-1 線形補間法（等間隔）の再現率はいずれも平均約 0.77 で精度が高く，(4)-2 線形補間法（時間間隔を考慮）の再現率は平均約 0.80 で最も精度が高かった. 補完後のデータを観察すると，欠測の直前・直後で異常の傾向が変わらないデータが多く，欠測の直前あるいは直後の値を用いる(1)前方補完法，(2)後方補完法，(4)線形補間法の精度が高くなったと考えられる. 一方で，欠測の前後だけではなく全体の値を考慮する(3)平均値代入法では，正確な値が補完できなかったと考えられる.

また，(3)平均値代入法，(4)線形補間法は時間間隔を考慮すると，(3)-3 に対して(3)-4 では約 18%，(4)-1 に対して(4)-2 では約 3%精度が上がった. 今回のデータセットには入院時だけでなく外来時の検査結果も含まれ，検査時間が年単位で異なる場合もあり，一定間隔ではない. そのため，時間間隔を考慮することにより補完精度が上がったと考えられる.

4. おわりに

本研究では，実際の電子カルテデータより欠測を含む血液検査結果を用いて，4 種類の補完法の適用可能割合と異常性に対する補完精度から，実用性を評価した. 適用可能割合は(3)-1 平均値代入法（患者全体）は全て 100%である一方，(4)線形補間法は平均約 76%で比較的低かった. 補完精度については，時間間隔を考慮することにより精度が向上した. 全体としては，(4)-2 線形補間法（時間間隔を考慮）の再現率が平均約 0.80 で最も精度が高かった. そこで今後は，補完精度が高い一方，適用可能割合が低く課題があった，線形補間法に注目し改善を試みる.

参考文献

[1] 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, 4C-1 (2025)

[2] Johnson,A. et al.:MIMIC-IV (version 3.1), PhysioNet (2024)

[3] Johnson,A. et al.:MIMIC-IV, a freely accessible electronic health record dataset, Scientific Data, 10, 1 (2023)

[4] Goldberger, A. et al.:PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation [Online], 101, 23, pp. e215-e220 (2000)

Copyright ©2026 Information Processing Society of Japan.
All Rights Reserved.