様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 ２８ 日現在

機関番号：２２７０１
研究種目：挑戦的研究（萌芽）
研究期間：2018～2023
課題番号：１８Ｋ１８４７１

研究課題名（和文）複数疾病を伴う高齢入院患者の予後予測因子の同定：機械学習モデルの解釈性の向上

研究課題名（英文）Identification of prognostic factors for elderly hospitalized patients with multiple diseases: improving interpretability of machine learning models.

研究代表者
清水　沙友里（Shimizu, Sayuri）

横浜市立大学・データサイエンス研究科・講師

研究者番号：６０６２５４０８

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：複数疾患を持つ身体的に脆弱な高齢者が増加していることにより、データベース研究におけてそれらを包括的な視点から評価を行い、臨床評価に役立てることは重要な課題であるが、複数疾病のある患者に対して、それらの並存パターンの重症度評価は十分ではない。本研究では、従来型モデルや勾配ブースティングモデル、解釈可能性を加味したモデルなど複数の手法による予測モデルを構築し、予測モデル精度の向上が可能であることが示唆された。本分析により、医療管理データの持つデータ特性と、機械学習モデルとの分析上の親和性を加味し、解釈可能性に留意しながら分析モデルを選択することの重要性が改めて示唆された。

研究成果の学術的意義や社会的意義
臨床現場から日々生成される医療データが蓄積され、世界的な潮流として、これらのデータを臨床や政策に活用しようという動きが広がっています。加えて、従来型の統計モデルから機械学習モデルへのシフトがおこっており、これらのモデルを医療管理分野の分析にどのように活かすかが課題となっていました。本研究では、機械学習モデルがより精度高く予測可能でありましたが、解釈可能性に留意する必要があることが示唆されました。

研究成果の概要（英文）：The increasing number of elderly individuals with multiple diseases and reduced physical resilience makes it imperative to evaluate them from a comprehensive perspective in database studies for clinical assessment. In this study, we constructed prediction models using multiple methods, including conventional models, gradient boosting models, and models that take interpretability into account, suggesting that it is possible to improve the accuracy of prediction models. This analysis reiterates the significance of selecting an analytical model that accounts for the distinctive characteristics of healthcare administrative data, the analytical compatibility with machine learning models, and interpretability.

研究分野：ヘルスサービスリサーチ

キーワード：医療データベース　機械学習

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

高齢化社会の到来を迎え、疾病構造の変化やマルチモビディティへの対応は世界各国の共通的政策課題の一つである。高齢になればなるほど、疾病状態からの機能回復は遅れがちであり、数多くの併存疾患を有し医学的複雑性が高く、社会経済的要因によって状態がより悪化し、また認知機能の低下などの心身医学的課題を抱えている複雑な症例は多い。このような状況下では、複合的な要因から患者アウトカムの悪化がみられることが多い。我が国は、世界に類を見ない超高齢社会に突入し、これらのフレイルな状態像の患者に対して、医療資源が限られる中で医療需要の増加に対応するという困難な課題に取り組まなくてはならない。

このような背景から、高齢者の脆弱性を包括的な視点から評価を行い、予後悪化の要因を明らかにするということは重要な課題である。とりわけ入院医療においては、在院日数を短縮化し効率化を図りながらも医療の質を担保していくことが求められている。医療機関毎に異なる患者背景や重症度を補正するため、Charlson comorbidity index 等の幾つかの患者重症度評価手法が用いられているが、診療実態に即した指標とはなっていない。加えて、地域の医療提供体制などの要因はデータ不足から充分に考慮されておらず、社会医学領域においては、予測力に劣る線形回帰モデルの利用から脱却できない故に複雑な疾病の併存状況のパターンを評価できていないなど、データ的・手法論的に挑戦可能な課題も数多く残っている。

２．研究の目的

本研究は、急性期医療機関に入院中の高齢者を対象として、①これまでに用いられてこなかった多様で詳細な医療大規模データの利用 ②XGBoost などの最新の機械学習手法の活用 ③機械学習手法を社会医学分野に応用するときに大きな課題となる‘ モデルのブラックボックス性 ’に対処する手法の導入 を行うことで、患者のマルチモビディティを考慮した、在院日数、退院時 ADL、自宅復帰、再入院率等の予後予測因子のより精緻な同定を行うことを目的とする。本研究は、喫緊の政策的課題である高齢者への持続的な医療提供体制の構築に向けた政策的課題への応用、機械学習手法の社会医学系データへの活用、そして人間の理解を超えた複雑な機械学習モデルに対する「説明力」の実装を通して、社会医学・機械学習両領域にとってブレイクスルーとなる予測モデルの開発が期待される。

３．研究の方法

プライマリーアウトカムは在院日数、１入院あたり出来高医療費とする。セカンダリーアウトカムは、院内死亡率、退院時 ADL、自宅復帰率、30 日以内再入院率とする。共変量として DB より患者居住地の医療資源整備状況、傷病名や疾患重症度等の医学的背景、病床数や医師看護師数などの医療機関特性を取得する。①一般的な予測モデルである多変量ロジスティック回帰モデル ②比較的古典的な手法である決定木モデル ③弱識別器を独立的に学習していく Random Forest ④弱識別器の学習を逐次的に行う Boosted Trees の一種である XGBoost という４種のモデル精度の比較を行う。Random Forest 及び XGBoost は多数の決定木を組み合わせるアプローチであり医学管理データのスパース性にも対応しながら、高い予測パフォーマンスが期待できるが、予測のメカニズムを人間が論理的に解釈することは困難で、いわば、これらのモデルは ’ ブラックボックス ’ である。本研究では、①複雑なモデルを解釈性の高いシンプルなモデルで記述し直す ②個々の症例に対するモデルによる予測結果を少数の要因で説明する 仕組みを導入し、ブラックボックスのホワイト化を試みる。

４．研究成果

解析結果

図１より、生存/死亡で入院日数に差があることから、入院日数の予測を行うことは妥当であると考えられたため、予測モデルの構築を行った。FF1 データ 326,024 件のうち、欠損データを除外した 135,275 件をデータ構築に用いた。図３より、適切な閾値を設定すれば、「入院日数が 3 日以下で死亡退院する人」が、「入院日数に限らず死亡退院する人」より高い精度で予測であるため、閾値を３日に設定した。病院の ID を one-hot ベクトル化して、予測の変数に加えて実験した時の AUC:0.851、病院 ID を全く予測に入れなかった時の AUC:0.847 となった。また、各モデル(ロジスティック回帰、ランダムフォレスト、勾配ブースティング決定木等)においての個々の AUC は精度が高く、機械学習モデルにおける LIME 及び shap 改良モデルによる説明可能性の向上は見られたが、臨床研究での利用のためには、ロジスティック回帰の汎用性と説明可能性モデルの理解の難しさについて考慮し、モデル利用を検討する必要があると示唆された。

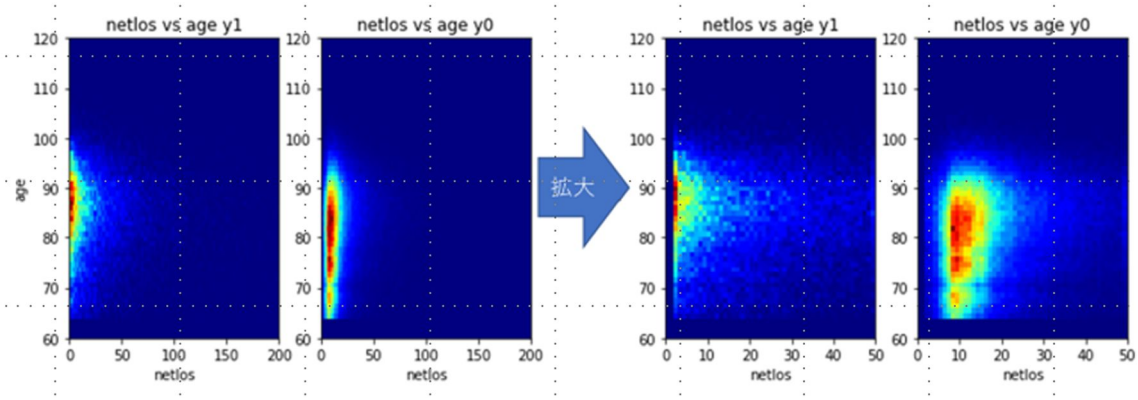

図 1.生存/死亡による入院日数の差

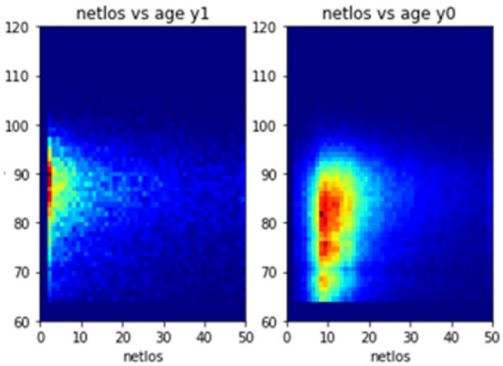

図 2.「入院日数が n 日以下」and「Y＝１(死亡)」のデータを新たに y=1 として、３日～１０日の間で変動

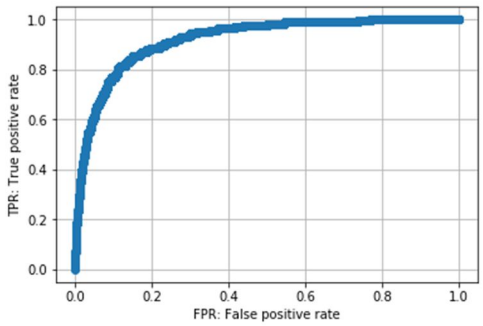

図 3．入院予測日数の閾値の設定

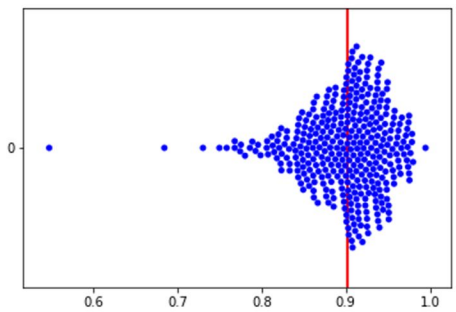

図 4．医療機関毎の死亡/生存の在院日数分布

thershold	rate	accuracy	auc
3	0.0082	0.9918	0.9212
4	0.0133	0.9867	0.9057
5	0.0180	0.9819	0.8964
6	0.0223	0.9776	0.8949
7	0.0263	0.9736	0.8891
8	0.0300	0.9701	0.8842
9	0.0332	0.9669	0.8787
10	0.0361	0.9640	0.8748

参考：データラベル(y)を生存/死亡のみにして、同条件で実験を行った場合
Rate:0.09776　accuracy:0.9059　auc:0.8470

図 4.分析モデルにおける精度

オープンデータの公開
本研究において作成した各年度の『全国保険医療機関一覧』、『二次医療圏―郵便番号一覧』においては、以下の research map の資料公開ページでダウンロード可能な形で公開している。

https://researchmap.jp/ssyr

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 清水沙友里	４．巻 6月号
２．論文標題 ビッグデータを対象とした解析の実際と読み方のポイント	５．発行年 2023年
３．雑誌名 Life Support and Anesthesia(LiSA)	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 清水沙友里	４．巻 71
２．論文標題 医療・健康分野におけるビッグデータ解析	５．発行年 2023年
３．雑誌名 会報光触媒	６．最初と最後の頁 41-46
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　2件／うち国際学会　1件）

１．発表者名 清水沙友里
２．発表標題 ビッグデータとAIにより広がる近未来予想図
３．学会等名 第33回日本臨床モニター学会総会（招待講演）
４．発表年 2022年

１．発表者名 清水沙友里
２．発表標題 生物統計セミナー「明日から使える医療統計 クリニカルクエスチョンから論文作成まで一気通貫 part 2」
３．学会等名 第264回日本循環器学会関東甲信越地方会（招待講演）
４．発表年 2022年

１．発表者名 Sayuri Shimizu, Satoshi Hara, Kiyohide Fushimi
２．発表標題 Predicting the risk of in-hospital Mortality in Adult Community-Acquired Pneumonia Patients with Machine Learning: A Retrospective Analysis of Routinely Collected Health Data
３．学会等名 ISPOR Europe 2019 conference（国際学会）
４．発表年 2019年

〔図書〕　計1件

１．著者名 後藤，隆久，原，広司，田中，利樹，黒木，淳，今中，雄一	４．発行年 2022年
２．出版社 中央経済社,中央経済グループパブリッシング	５．総ページ数 286
３．書名 データで変える病院経営　　第８章　ビッグデータを活用する	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	原　聡 (Hara Satoshi) (40780721)	大阪大学・産業科学研究所・准教授 (14401)	
研究分担者	伏見　清秀 (Fushimi Kiyohide) (50270913)	東京医科歯科大学・大学院医歯学総合研究科・教授 (12602)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関