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　５ 年　 ６ 月　 ８ 日現在

機関番号：１２１０２
研究種目：若手研究
研究期間：2019～2022
課題番号：１９Ｋ１９４３０

研究課題名（和文）大規模入院データを用いた機械学習による再入院予測モデルの構築

研究課題名（英文）Development of risk prediction model for re-admission in large inpatient data with machine learning

研究代表者

岩上　将夫（Iwagami, Masao）

筑波大学・医学医療系・准教授

研究者番号：３０８３０２２８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：大規模入院データ(Medical Data Vision社から購入した38病院のDPC情報および採血検査結果)から30日以内の予定外再入院を予測する複数の機械学習・ロジスティック回帰モデルを構築し、その予測能の比較を行った。その結果、機械学習の一種であるgradient-boosted decision tree(GBDT)が最も識別能に優れることが明らかになった。一方で、予測変数の項目数を増やし採血結果を予測に用いるほど、機械学習のベネフィットが高まることを期待していたが、最も項目数の多い(採血結果含む1543項目)データセットにおいて、GBDTとロジスティック回帰の識別能に大差はなかった。

研究成果の学術的意義や社会的意義
近年リアルワールドデータ(電子カルテや医療レセプト)の収集・利活用や、大規模医療データに対して機械学習を適用することへの期待が高まっている。そこで本研究では、日本の退院患者の再入院予測を例に、機械学習と昔から使われていたロジスティック回帰モデルの予測能を比較する実験を行った。その結果、確かに機械学習の一種であるgradient-boosted decision tree(GBDT)が最も判別能に優れていたが、一方で、必ずしも多くの情報量を利用する時ほどそのベネフィットが高まるわけではないことも明らかになった。以上の結果は、今後日本で医療情報に機械学習を適用し社会実装する際に参考になるであろう。

研究成果の概要（英文）：We compared the predictive performance of gradient-boosted decision tree (GBDT), random forest (RF), deep neural network (DNN), and logistic regression with the least absolute shrinkage and selection operator (LR-LASSO) for 30-day unplanned readmission. We used electronic health records of patients discharged alive from 38 hospitals. We created six patterns of datasets having different numbers of binary variables (that over 5% or 1% of patients or 10 patients had) with and without blood-test results. For the dataset with the smallest number of variables (102), the c-statistic was highest for GBDT (0.740), followed by RF (0.734), LR-LASSO (0.720), and DNN (0.664). For the dataset with the largest number of variables (1543), the c-statistic was highest for GBDT (0.764), followed by LR-LASSO (0.755), RF (0.751), and DNN (0.720). We found that GBDT generally outperformed LR-LASSO, but the difference became smaller as the number of variables was increased and blood-test results were used.

研究分野：疫学・予防医学

キーワード：機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

予定外再入院は、医療費、医療スタッフ、患者に負担を強いるため、公衆衛生上の課題である。予定外再入院の中には、予防可能なものもあるため、そのリスクが高い患者を特定し、介入することが重要である。これを実現するためには、個々の患者の再入院の確率を正確に推定できる臨床予測モデルが不可欠である。これまで予定外再入院に対する臨床予測モデルは、少数の臨床的変数に基づくロジスティック回帰(LR)モデルを中心に開発されてきた。しかし、これらのモデルの予測能は、C 統計(ROC 曲線下面積)が 0.70 未満と、良いものとは言えない。予測能力は、予測変数の数または種類を増やし、また最新のより正確な統計的または数学的モデルを開発することで向上させるべきと考えられている。

最近では、勾配ブースト決定木(GBDT)、ランダムフォレスト(RF)、ディープニューラルネットワーク(DNN)などの機械学習モデルが、リアルワルドデータ(電子カルテやレセプトデータ)を用いた再入院予測モデルの構築に適用されている。しかし、最近のシステマティックレビュー(J Clin Epidemiol. 2019,110,12-22)によると、伝統的な回帰モデル(LR や Cox 回帰モデル)と機械学習モデルの間の予測能に統計的有意差はなく、平均 C 統計値はそれぞれ 26 研究で 0.71、15 研究で 0.74、その差は 0.03 (95%信頼区間[CI]: -0.01～0.07)であった。しかし、その大きな限界として、システマティックレビューに含まれた研究のうち、同じ研究内で従来の回帰モデルと機械学習モデルの直接比較を行ったものはわずかであった。さらに、システマティックレビューに含まれた研究には、予測変数の数が少ない研究が多かった。機械学習モデルの予測能は、予測変数の数だけでなく、血液検査結果のようなアウトカムと非線形な関係を持つ可能性のある連続変数を含めるかどうかにも依存する可能性がある。つまり、機械学習モデルの中には、予測変数の数を増やし、血液検査結果を予測に用いることで、LR モデルの予測能を上回るモデルが構築できる可能性も期待できる。

２．研究の目的

そこで本研究では、日本の 38 病院から入手した電子カルテ情報を用い、予測変数の数(患者の 5%以上、1%以上、10 人以上が有する変数) および血液検査結果の有無で 6 パターンのデータセットを意図的に作成し、30 日以内の予定外再入院予測のための機械学習(GBDT、RF、DNN)および least absolute shrinkage and selection operator(LASSO)を用いた LR モデル (LR-LASSO) を開発・検証・比較することを目的とした（注：LR の特徴量選択方法として、ステップワイズ変数選択や LASSO などがあるが、一般的に LASSO は他の手法に比べて過学習を起こしにくいこと、および再入院の先行研究で LR-LASSO がステップワイズ変数選択を用いた LR よりも予測能が優れていたことから、本研究では LR-LASSO を用いることとした）。

３．研究の方法

(1)データソース

本研究は、Medical Data Vision(MDV)データベースを用いた後ろ向きコホート研究である。MDV データベースは、メディカル・データ・ビジョン社が構築した匿名加工情報で、日本の Diagnosis Procedure Combination(DPC)制度参加病院の中で、MDV 社の業務支援システムを利用し、研究目的での二次データ利用に同意した、350 以上の急性期病院から構成されている。本研究では、さらに採血検査値のデータの提供に同意した 38 病院から取得したデータを用いて研究を行った。

DPC データには、年齢や性別などの基本的な患者特性、退院時に担当医が記録した診断名(入院病名、主病名、最も/2 番目に資源を消費した病名、入院時併存疾患、入院後合併症の ICD-10 コード)、手術コード、人工呼吸や透析などの処置コード、処方薬コード(ATC コード)、入院状況(予定入院か予定外か)、退院状況(死亡、施設や自宅への退院、他病院への転院)がわかる。

本研究は、ヘルシンキ宣言に基づき、筑波大学倫理委員会の承認（承認番号 1414）を得た。

(2)研究対象者

2015 年 1 月 1 日から 2018 年 12 月 31 日までに、38 病院に入院し、入院中に少なくとも 1 回の血液検査を行った患者の退院を同定した。(i)入院病名カテゴリーが出産の症例(ICD-10 の O00-Q99)、(ii)死亡退院、(iii)他院への転院、(iv)退院先または 10 項目の基本血液検査(白血球数、ヘモグロビン、血小板数、ナトリウム、カリウム、塩素、クレアチニン、BUN、AST、ALT)が 1 項目以上欠損しているもの、を除外した。同一患者が研究期間中に複数回入院した場合、各入院を独立した入院とみなした。2015 年 1 月 1 日から 2017 年 12 月 31 日に退院した患者のデータを用いてモデルの開発を行い、2018 年 1 月 1 日から 12 月 31 日に退院した患者のデータを用いて予測能を検証した。

(3) アウトカム

患者が退院した同病院への 30 日以内の予定外再入院とした。

(4) 予測変数

本研究で使用した予測変数は、基本特性（年齢、性別、入院診断カテゴリー、過去 1 年間の入院回数、退院先)、記録された診断病名(3 桁の ICD-10 コード合計 2102 種類の有無)、入院中の治療（914 種類の手術コード、122 種類の処置コード、468 種類の ATC コードの有無)、退院前の最

後の血液検査結果から構成されている。完全分離(perfect separation)の問題を避けるため、10人未満の患者しか有していない変数を除外した。次に、以上の二値予測変数の候補の中から、異なる頻度の変数(患者の 5%以上または 1%以上または 10 名以上が有する変数)および血液検査結果の利用の有無により 6 パターンのデータセットを意図的に作成した.

(5)統計解析

まず、開発データと検証データにおける、アウトカムの有無に応じて、すべての予測変数の分布または比率を記述した。次に、上記の 6 パターンの各データセットについて、開発データを用いて機械学習モデル(GBDT、RF、または DNN)および LR-LASSO モデルを構築した。各モデルのハイパーパラメーターは、統計ソフト R の h2o パッケージによる自動機械学習により、10 分割クロスバリデーションにより決定・最適化した。次に、検証データを用いて、各モデルの性能を c 統計量(ROC 曲線下面積)およびキャリブレーションプロットで評価した。機械学習モデルと LR-LASSO の c 統計量を比較するため、Delong の検定を実施した。最後に、検討したモデルの中で最高の性能を示した GBDT に着目し、重要度の高い変数を特定した。

統計ソフトとして、データクリーニングは STATA version 16 を用いて行い、統計解析は R version 4.1.2 を用いて h2o パッケージと rms パッケージを使用して行った。

４．研究成果

38 病院 410,941 人の退院 635,509 件から、上述の除外基準を適用した結果、457,587 件が解析対象となり、開発データセットには 339,513 件(平均年齢 62.0±24.6 歳、男性 54.3%)、検証データセットでは 118,074 退院(平均年齢 63.4±24.1 歳、男性 54.1%)が含まれた。30 日以内の予定外再入院の発生率は、それぞれ 6.8%(23,108/339,513)、6.4%(7,507/118,074)であった。

上述の 6 パターンのデータセットを作成した結果、パターン 1（5%以上の患者が有していた二値変数、採血検査結果は利用しない）には 102 変数、パターン 2(5%以上の患者が有していた二値変数および採血検査結果）には 112 変数、パターン 3（1%以上の患者が有していた二値変数、採血検査結果は利用しない）には 296 変数、 パターン 4（1%以上の患者が有していた二値変数および採血検査結果）には 306 変数、パターン 5（10 以上の患者が有していた二値変数、採血検査結果は利用しない）には 1533 変数、パターン 6（10 以上の患者が有していた二値変数および採血検査結果）には 1543 変数が含まれた。

各パターンのデータセットにつき、各機械学習・LR モデルの C 統計量を表に示す。GBDT はすべてのパターンで最も C 統計量が高く、LR-LASSO を統計学的に有意に上回った。ただし、変数の数が増え、また採血結果を利用するほど、GBDT がより予測に優れるというわけではなく、パターン 6（10 以上の患者が有していた二値変数および採血検査結果）では GBDT と LR-LASSO の C 統計量（それぞれ 0.764、0.755）の差がむしろ小さくなることが明らかになった。キャリブレーションについては、再入院のリスクが低い患者については、どのモデルも同様にキャリブレーションが良好であった。一方、再入院のリスクが高い患者については、GBDT と LR-LASSO は再入院確率を過大評価する傾向があり、RF と DNN は再入院の過小評価する傾向が見られた。

なお、最も識別能が高かった GBDT の feature importance の上位には、年齢、血液検査結果、過去 1 年間の入院回数が含まれ、これらが重要な再入院予測因子であることが示唆された。

表：各モデルの C 統計量(点推定値、95%信頼区間、LR-LASSO と比較した P 値)

データセット	GBDT	RF	DNN	LR-LASSO
パターン 1	0.740 (0.735 – 0.746) P<0.001	0.734 (0.729 – 0.740) P<0.001	0.664 (0.658 – 0.670) P<0.001	0.720 (0.714 – 0.726) n/a
パターン 2	0.751 (0.745 – 0.756) P<0.001	0.742 (0.736 – 0.747) P<0.001	0.728 (0.722 – 0.734) P<0.001	0.734 (0.728 – 0.740) n/a
パターン 3	0.756 (0.751 – 0.762) P<0.001	0.747 (0.742 – 0.753) P<0.001	0.692 (0.685 – 0.698) P<0.001	0.740 (0.734 – 0.746) n/a
パターン 4	0.759 (0.754 – 0.765) P<0.001	0.753 (0.748 – 0.759) P=0.013	0.737 (0.731 – 0.743) P<0.001	0.749 (0.744 – 0.755) n/a
パターン 5	0.759 (0.753 – 0.764) P<0.001	0.747 (0.741 – 0.752) P=0.612	0.701 (0.695 – 0.707) P<0.001	0.747 (0.742 – 0.753) n/a
パターン 6	0.764 (0.758 – 0.769) P<0.001	0.751 (0.746 – 0.757) P=0.010	0.720 (0.714 – 0.726) P<0.001	0.755 (0.749 – 0.761) n/a

以上から、日本の大規模電子カルテデータを用いて 30 日以内の予定外再入院を予測する際には、機械学習の中でも GBDT が従来の LR モデルよりも予測能に優れることが示された。ただし、仮説に反して、必ずしも予測変数の数が増え、採血結果のような連続変数が含まれたからといって、機械学習のベネフィットが高まるわけではないことも発見した。これらの知見は、今後日本で医療情報に機械学習を適用し社会実装する際に参考になることが期待できる。

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　2件）

1．著者名 Masao Iwagami, Ryota Inokuchi, Eiryo Kawakami, Tomohide Yamada, Atsushi Goto, Toshiki Kuno, Yohei Hashimoto, Nobuaki Michihata, Tadahiro Goto, Tomohiro Shinozaki, Yu Sun, Yuta Taniguchi, Jun Komiyama, Kazuaki Uda, Toshikazu Abe, Nanako Tamiya	4．巻 -
2．論文標題 Comparison of machine-learning and logistic regression models to predict 30-day unplanned readmission: a development and validation study	5．発行年 2023年
3．雑誌名 medRxiv	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） 10.1101/2023.05.06.23289569	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Masao Iwagami, Hiroki Matsui	4．巻 4(3)
2．論文標題 Introduction to Clinical Prediction Models	5．発行年 2022年
3．雑誌名 Annal of Clinical Epidemiology	6．最初と最後の頁 72-80
掲載論文のDOI（デジタルオブジェクト識別子） 10.37737/ace.22010	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 岩上将夫、井口竜太、川上英良、山田朋英、後藤温、橋本洋平、道端伸明、小宮山潤、宇田和晃、田宮菜奈子
2．発表標題 大規模入院データによる再入院予測のための複数の機械学習法の比較
3．学会等名 日本臨床疫学会第5回年次学術大会
4．発表年 2022年

〔図書〕　計0件

〔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

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