氏　　名　伊庭 克拓

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2317 号

学位授与の日付　2022 年 3 月 24 日

学位授与の要件　複合科学研究科 統計科学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　多変量臨床予測モデルにおけるリサンプリング法に基づく内的検証法の評価研究

論 文 審 査 委 員　主　査　間野 修平
統計科学専攻 教授
野間 久史
統計科学専攻 准教授
矢野 恵佑
統計科学専攻 准教授
菅澤 翔之助
東京大学 空間情報科学研究センター 准教授

（様式３）

博士論文の要旨

氏　　名　伊庭　克拓

論文題目　多変量臨床予測モデルにおけるリサンプリング法に基づく内的検証法の評価研究

背景と目的

多変量臨床予測モデルは、患者の複数の特性に基づいて、診断および予後の予測を行うための重要な統計ツールの１つである。予測モデルの構築に用いたデータで評価したモデルの判別・較正などの予測精度の指標は、オプティミスム（optimism）と呼ばれる過大評価のバイアスを含んでおり、将来予測を行う外部集団での予測精度よりも過大に推定されることが知られている。多変量予測モデルに関するガイドラインである TRIPOD 声明では、ブートストラップなどのリサンプリング法を用いた内的検証法によるオプティミスムの調整を推奨している。

ブートストラップによる内的検証法として、Harrell 法、Efron の.632 法及び.632+法が提案されている。現在、Harrell 法が慣例的に大半の研究で使われているが、元来、Efron の.632 法及び.632+法は、Harrell 法のような単純なバイアス補正法を改良するために開発された手法であり、より正確な推定値を得られることが期待できる。しかしながら、これまでに、これらの手法の有用性を、実践的な条件下で詳細に評価した研究はわずかしかなく、限定的なエビデンスしか得られていない。特に、過去の研究では、考慮されている条件が限定的であり、近年、臨床研究の実践においても普及しつつある正則化法及び変数選択のために現在も用いられているステップワイズ法を含めた比較・評価は行われていなかった。

そのため、広範な実践的条件のもとで、これらのリサンプリング法に基づく内的検証法の性能を比較・評価し、臨床研究の実践における新規なガイドラインを与えることを目的として、シミュレーション実験を行った。特に、従来のロジスティック回帰（最尤法）に加え、ステップワイズ法などの変数選択法、また、Firth 法、Ridge 回帰、Lasso 回帰及び Elastic-net 回帰など、最新のモデル構築法を用いた条件下での性能評価まで、詳細な分析を行った。

また、標準的な内的検証法であるブートストラップ法について、これまで信頼区間の補正法は提案されていなかったため、オプティミスムを補正した推定量に基づく信頼区間（位置補正（location-shifted）ブートストラップ法及び２段階（two-stage）ブートストラップ法）の開発を行った。

方法

急性心筋梗塞の治療法の有効性を評価した欧米での大規模ランダム化臨床試験である GUSTO-I 試験のデータセットに基づいた広範な設定でシミュレーション実験を行った。イベント変数は心筋梗塞の発症後 30 日の死亡であり、17 個の予測変数のデータが得られている。

予測精度に影響する可能性がある要因として、予測変数あたりのイベント数（EPV）、イベント発生割合、候補の予測変数の数及び予測変数の回帰係数を考慮した。予測変数のデータは、GUSTO-I データセットから推定したパラメータを基に生成した。結果変数は、これらの予測変数に基づきロジスティック回帰モデルから生成した。シミュレーション及びブートストラップリサンプリングの回数は 2000 回に設定した。7 つのモデル構築法（最尤法、Firth 法、Ridge 回帰、Lasso 回帰、Elastic-net 回帰及び backward ステップワイズ法（AIC 及び $P < 0.05$））によって、予測モデルを構築した。モデル構築に用いたデータセットに対する未調整の C 統計量、Harrell 法、Efron の.632 法及び.632+法によるオプティミスムを調整した C 統計量を求めた。500,000 例の検証データセットに対する外部の C 統計量を真値とみなして、未調整及び各内的検証法の C 統計量のバイアス及び RMSE（root mean squared error）を評価した。

提案した信頼区間の妥当性を確認するために、従来の未調整の信頼区間を比較対象として、同様のシミュレーション実験によって、被覆確率と信頼区間幅を評価した。

結果

一定以上の規模のサンプルサイズ（EPV が 10 以上）のもとでは、いずれのブートストラップ法に基づく内的検証法の C 統計量にもバイアスがなく、概ね妥当な推定値が得られた。小標本下では、Harrell 法及び.632 法は同様の傾向であり、イベントの発生割合が大きい場合に過大評価のバイアスを示した。また、イベントの発生割合が小さい場合、.632+法は若干の過小評価のバイアスを示す傾向があった。他の２つの方法と比較して、.632+法のバイアスは相対的に小さかったが、RMSE は同程度もしくは特に小標本のもとで正則化法が用いられた場合においては大きい傾向が認められた。

未調整の信頼区間の被覆確率は名義水準を大きく下回っていた。位置補正ブートストラップ法は、一定以上の規模のサンプルサイズでは被覆確率が概ね名義水準であり、２段階ブートストラップ法は、ほとんどの条件で被覆確率は名義水準付近であった。

考察

リサンプリング法に基づく内的検証法の性能を、広範な設定のもとで、大規模なシミュレーション実験により評価した。比較的サンプルサイズの大きな条件のもとでは、３つのブートストラップ推定量の性能は、概ね同等であり、いずれもほとんどバイアスは認められなかった。一方、小標本のもとでは、３つの推定量にはいずれにもバイアスがあり、バイアスの方向と大きさには一貫性がなかったが、正則化法が用いられた場合にばらつきが大きくなる点を除いて、.632+推定量の性能が相対的に優れていた。したがって、一般的には、これまで慣例的に用いられてきた Harrell 法よりも、.632+法の使用が推奨される。ただし、小標本のもとで正則化法が用いられる条件下では、ばらつきが大きくなることに注意する必要がある。

従来の未調整の信頼区間は、現実的な条件下では被覆確率が名義水準を大幅に下回っており、実践おいては推奨されない。提案する信頼区間によって、より高い正確性で、予測精度の指標の区間推定を行うことが可能になった。

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏　　名　伊庭　克拓

Title
論文題目　多変量臨床予測モデルにおけるリサンプリング法に基づく内的検証法の評価研究

【論文の概要】

提出された論文は，臨床医学研究において，疾患の診断や予後予測のための多変量臨床予測モデルの開発に用いられるリサンプリング法に基づく内的検証法の比較有用性，および，リサンプリング法に基づく予測精度の指標の信頼区間の構成方法について論じたものである．和文で書かれており，全５章による計１０２頁からなる．

第１章は，本論文の序章となっており，医学研究における多変量臨床予測モデルの開発に関連する既存研究やガイドラインなど，本研究の学術的背景について述べられている．第２章では，本研究で用いられる多変量予測モデルとその推測手法について解説がなされている．第３章で，リサンプリング法に基づく内的検証法の比較有用性の評価研究について述べられている．まず，予測性能の評価に用いられる判別・較正指標と，モデルの過剰適合が原因となって起こる過大評価のバイアス，および，リサンプリング法に基づくその内的検証法として，Harrell のバイアス補正法，0.632 法，0.632+法について解説されている．次に，欧米で行われた，急性心筋梗塞の大規模ランダム化臨床試験である GUSTO-I 試験の事例解析が示されている．その後，スーパーコンピュータを駆使した大規模シミュレーション実験による，これらの内的検証法の比較有用性についての分析について述べられている．GUSTO-I 試験をモデルとした，広範な実践的条件のもとで実験は行われており，標準的な最尤法，Ridge 回帰，Lasso 回帰，Elastic-net 回帰といった正則化法，変数選択に用いられるステップワイズ法などのさまざまなモデル構築法のもとでの詳細な分析が行われている．性能評価には，判別指標として，実践上，最もよく用いられている C 統計量（ROC 曲線の曲線下面積の推定量に対応する指標）のバイアスと RMSE（root mean squared error）が用いられている．シミュレーション実験の結果を総括すると，一定以上の規模のサンプルサイズ（event-per-variable が 10 以上）のもとでは，いずれの内的検証法の C 統計量にもバイアスはなく，概ね妥当な推定値が得られるという結果が一貫して得られていた．それよりもサンプルサイズが小さい規模の条件下では，Harrell のバイアス補正法と 0.632 法は，イベントの発生割合が大きい場合に過大評価のバイアスを示す傾向があった．一方，0.632+法は，イベントの発生割合が小さい場合に，若干の過小評価のバイアスを示す傾向があった．これらの条件下でも，相対的なバイアスは 0.632+法が最も小さく，また保守的な評価を与えるという傾向があったが，RMSE は概ね同程度であった．一方，正則化法が用いられた場合においては，0.632+法の RMSE が大きくなる傾向が認められた．第４章では，リサンプリング法に基づく予測精度の指標の信頼区間の構成方法について述べられている．提案された方法は，位置補正ブートストラップ信頼区間，２段階ブートストラップ信頼区

間の２つである．前者は，予測精度の指標の単純なブートストラップ信頼区間を，リサンプリング法に基づく方法から推定されるバイアスに基づいて位置補正するという方法となっている．後者は，対象集団からの２段階のリサンプリングを行い，バイアス補正後の推定量のブートストラップ分布を直接求め，そこから信頼限界を算出するという方法となっている．GUSTO-I 試験を事例として，その有用性は評価されており，また，スーパーコンピュータを駆使した大規模シミュレーション実験によって，従来の方法よりも格段に高い被覆率を達成することが示されている．第５章で，本研究についての考察，および，今後の課題がまとめられている．

【論文の評価】

多変量予測モデルによる統計解析は，臨床医学研究における疾患の診断と予後予測において，複数の臨床変数の情報を活用することができる有用な方法であり，近年では，機械学習の手法を用いた高度な分析による AI 診断機器等の開発もさかんに行われている．2015 年には，多変量予測モデルを用いた臨床医学研究のガイドラインとして TRIPOD 声明も公表されている．本研究が対象とするリサンプリング法に基づく内的検証法は，TRIPOD 声明による推奨を受け，過去 5 年ほどで，臨床医学研究の実践において急速に普及した方法である．しかしながら，現状では，ほとんどの研究において，最も単純な Harrell のバイアス補正法が，明確な科学的根拠のないまま，慣例的に用いられており，その他の方法の実践上の有用性について詳細な分析が行われた研究も，これまでにほとんどなかった．加えて，その信頼区間の構成におけるバイアス補正の必要性について論じられた研究もこれまでに皆無であった．予測精度の指標のバイアス補正の正確性・精度の分析は，開発される予測モデルの性能の評価に直接的に関わる要因であり，最適な方法が用いられないことで，医療現場で用いられる診断・予後予測，また，それに伴う意思決定が不適切なものになる可能性がある．本研究では，スーパーコンピュータを駆使した，これまでにない規模での大規模シミュレーション実験が行われており，さまざまな条件および手法を用いたもとでの主要な内的検証法の有用性を詳細に比較・分析し，また，新たな信頼区間の構成方法が提案されている．特に，昨今の機械学習の振興によって臨床医学研究の実践でも普及しつつある Lasso 回帰，Elastic-net 回帰といった正則化法を含めた分析は，計算コストの問題から，これまでこのような大規模な評価研究はほとんど行われてこなかったが，本研究では，スーパーコンピュータを駆使した大規模計算によって，その詳細な分析を可能としている．

本研究では，大規模な計算機実験の結果，臨床医学研究の実践において，これまでにほとんど用いられてこなかった 0.632+法が，広範な条件下で最も推奨される方法であるという新たな知見が得られている．また，予測精度の指標の信頼区間については，現在のスタンダードとなっている方法が誤ったものであることが明確に示されており，世界的にも初めてとなる妥当な信頼区間の構成方法の開発に成功している．本研究で得られた知見は，これまでの理論・実証研究からは得られなかった新規な知見であり，臨床医学研究の実践を変え得る重要なエビデンスを与えたものと評価することができる．以上をもって，審査委員会では，本論文が博士（統計科学）の学位を授与するに十分な水準を達成するものであると判定した．

【その他】

本研究の内容をまとめた研究論文が，査読付き国際学術誌 BMC Medical Research Methodology 誌，Statistics in Medicine 誌に掲載されている．