2021年6月18日　統計数理研究所　オープンハウス

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

伊庭　克拓　　総合研究大学院大学 統計科学専攻 博士課程5年

背景

複数の予測変数に基づいて疾患の診断及び予後の予測を行うために，多変量臨床予測モデルが用いられている．多変量予測モデルは，2値のアウトカムに対するロジスティック回帰モデルなど，適切な回帰モデルに基づいて構築される．モデルの構築に用いたデータで評価したモデルの判別・較正などの予測能力は，将来予測を行う外部集団での予測能力よりも高くなることが知られており，この過大評価のバイアスはオプティミスム（Optimism）と呼ばれている．多変量予測モデルの開発及び報告に関するガイドラインであるTRIPOD声明は，ブートストラップなどのリサンプリング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．代表的なブートストラップによるオプティミスムの調整方法として，Harrell et al. (1996) のバイアス補正法，Efronの.632法及び.632+法(Efron, 1983; Efron and Tibshirani, 1997) が提案されている．現在，比較的にシンプルなアルゴリズムで実行できるHarrell法が，慣例的によく使われている．一方，Efronの.632法及び.632+法は先行研究においてHarrell法と同等の性能であると報告されており，あまり使用されていない．これらの推定量は異なるコンセプトに基づいていることから，実践的な状況において異なる性能を示す可能性があるが，臨床研究の実践におけるこれらの方法の使用に関するガイダンスを出せるだけのエビデンスは十分に確立されていない．

本研究では，リサンプリング法に基づく内的検証法の性能を比較・評価することにより，臨床研究の実践におけるガイドラインを与えることを目的として，10未満のEPV（イベントあたりの予測変数の数）といった，これまでに十分評価されていない条件下でのエビデンスを与えるために，広範な設定でシミュレーションによる評価を行った．特に，従来のロジスティック回帰(ML法)，stepwise法，Firth法，ridge，lasso及びelastic-netなど，最新の多変量予測モデル構築方法を用いた場合の性能まで詳細に評価した．

Harrellのバイアス補正法

オプティミスムを調整するHarrellのバイアス補正法のアルゴリズムは，以下の通りである．

- オリジナル集団における未調整の予測能力の指標の推定値をθ_{app}とする．
- オリジナル集団からのリサンプリングによって，B組のブートストラップ標本を生成する．
- それぞれのブートストラップ標本に対して予測モデルを構築し，その予測能力の指標の推定値$\theta_{1,boot}, \theta_{2,boot}, \cdots, \theta_{B,boot}$を求める．
- ブートストラップ標本から構築されたB個の予測モデルを用いて，オリジナル集団に対する予測能力の指標の推定値$\theta_{1,orig}, \theta_{2,orig}, \cdots, \theta_{B,orig}$を求める．

オプティミスムのブートストラップ推定値は，以下となる．

$$\Lambda = \frac{1}{B}\sum_{b=1}^{B}\left(\theta_{b,boot} - \theta_{b,orig}\right)$$

未調整の予測能力の指標の推定値からオプティミスムの推定値を差し引くことにより，バイアスを補正した予測能力の指標の推定値$\theta_{app} - \Lambda$を得る．

ブートストラップ標本には，平均的にオリジナル集団の63.2%のデータが含まれるため，ブートストラップ標本とオリジナル集団にデータのオーバーラップが生じていることから，Harrellのバイアス補正法は，バイアスを過小評価する可能性がある．

Efronの.632法

Efronの.632法は，B組のブートストラップ標本を生成し，それぞれのブートストラップ標本に対して予測モデルを構築するところまでは，Harrellのバイアス補正法と同じである．.632法では，B組のブートストラップ標本について，ブートストラップ標本に含まれなかった外部標本を，B個の予測モデルに対するテストデータセットとみなし，ブートストラップ標本から構築されたB個の予測モデルを用いて，外部標本に対する予測能力の指標の推定値$\theta_{1,out}, \theta_{2,out}, \cdots, \theta_{B,out}$を求める．Efronの.632推定量は，未調整の予測能力の指標の推定値と外部標本に対する予測能力の指標の推定値の重み付き平均で構成される．

$$\theta_{.632} = 0.368 \times \theta_{app} + 0.632 \times \theta_{out}$$

$$\theta_{out} = \sum_{b=1}^{B} \theta_{b,out}/B$$

しかしながら，.632法はオーバーフィッティングの程度が強いモデルの場合に，オプティミスムを補正しきれないことが知られている．

Efronの.632+法

Efronの.632+法は，オーバーフィッティングの程度を考慮するために，オーバーフィッティング率

$$R = \frac{\theta_{out} - \theta_{app}}{\gamma - \theta_{app}}$$

を用いた.632法の改良版である．γは，無情報モデルの予測能力の指標の推定値である（例えば，C統計量の場合の理論値は0.5）．オーバーフィッティング率は，オーバーフィッティングがない（$\theta_{out} = \theta_{app}$）とき，0に近づき，オーバーフィッティングの度合いが強いとき，1に近づく．Efronの.632+推定量は，以下で定義される．

$$\theta_{.632+} = (1-w) \times \theta_{app} + w \times \theta_{out}$$

$$w = \frac{.632}{1 - .368 \times R}$$

Efronの.632+推定量は，オーバーフィッティングがないとき，.632推定量に近づき，オーバーフィッティングの度合いが強いとき，外部集団における推定値θ_{out}に近づく．

シミュレーション研究

本研究では，実際の臨床データであるGUSTO-I試験(The GUSTO Investigators, 1993)の一部であるGUSTO-I Western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を行った．GUSTO-I試験は，急性心筋梗塞のための4つの治療ストラテジーを評価した大規模臨床試験であり，多変量予測モデルの研究でも頻繁に用いられている．イベント変数は30日後の死亡であり，17個の予測変数が観測されている．

多変量予測モデルの判別性能の指標として最もよく用いられているC統計量を評価に用いた．C統計量は，イベント発生確率の推定値によるROC曲線のAUCのノンパラメトリックな推定量である．

予測能力に影響する要因として，イベントあたりの予測変数の数（EPV=3，5，10，20及び40），イベントの発生割合（0.5，0.25，0.125及び0.0625），候補の予測変数の数（先行研究で用いられた8変数及び全17変数）及び予測変数の効果（2シナリオ）を変動させ，合計80の設定で検討を行った．予測変数の効果（切片以外の回帰係数の真値）は，シナリオ1ではGUSTO-I Westernデータセットに対するML法の推定値を設定し，シナリオ2では予測変数の影響が小さい又はいくつかの予測変数が寄与しないことを仮定し，elastic-netの縮小推定値を設定した．切片の真値は，イベントの発生割合を調整するために設定した．

予測変数は，GUSTO-I Westernデータセットから推定したパラメータを基に，連続量は多変量正規分布，順序変数は多項分布，2値変数は多変量二項分布に従う乱数で発生させた．イベント発生確率π_iは，予測変数$\boldsymbol{x}_i$からロジスティックモデル$\pi_i = 1/\{1 + \exp(-\boldsymbol{\beta}'\boldsymbol{x}_i)\}$に基づいて決定した．イベント変数$y_i$は，ベルヌーイ分布$Bernoulli(\pi_i)$からの乱数で発生させた．

外部集団に対する予測性能は，独立に発生させた50万例のテストデータで評価した．シミュレーションの反復回数は2000とし，各反復でN（候補の予測変数の数×EPV/イベントの割合）例の学習データを発生させた．発生させた学習データに基づき，各モデル構築法(ML法，Firth法，ridge，lasso，elastic-net，stepwise法(p<0.05及びAICに基づく)によって多変量予測モデルを構築した．学習データに対する未調整のC統計量及び2000回のブートストラップリサンプリングを行ってHarrell法，Efronの.632法及び.632+法によるOptimismを調整したC統計量を求めた．テストデータに対するC統計量を真値とし，未調整及び各内的検証法のC統計量について，バイアスとRMSE(root mean squared error)を評価した．

結果

予測変数の効果がシナリオ2で，イベントの発生割合0.5の場合のバイアスの結果を図1に示した．ある程度の標本サイズ（EPVが10以上）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，上手く機能した．小標本では，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正則化法が用いられた場合において大きかった．

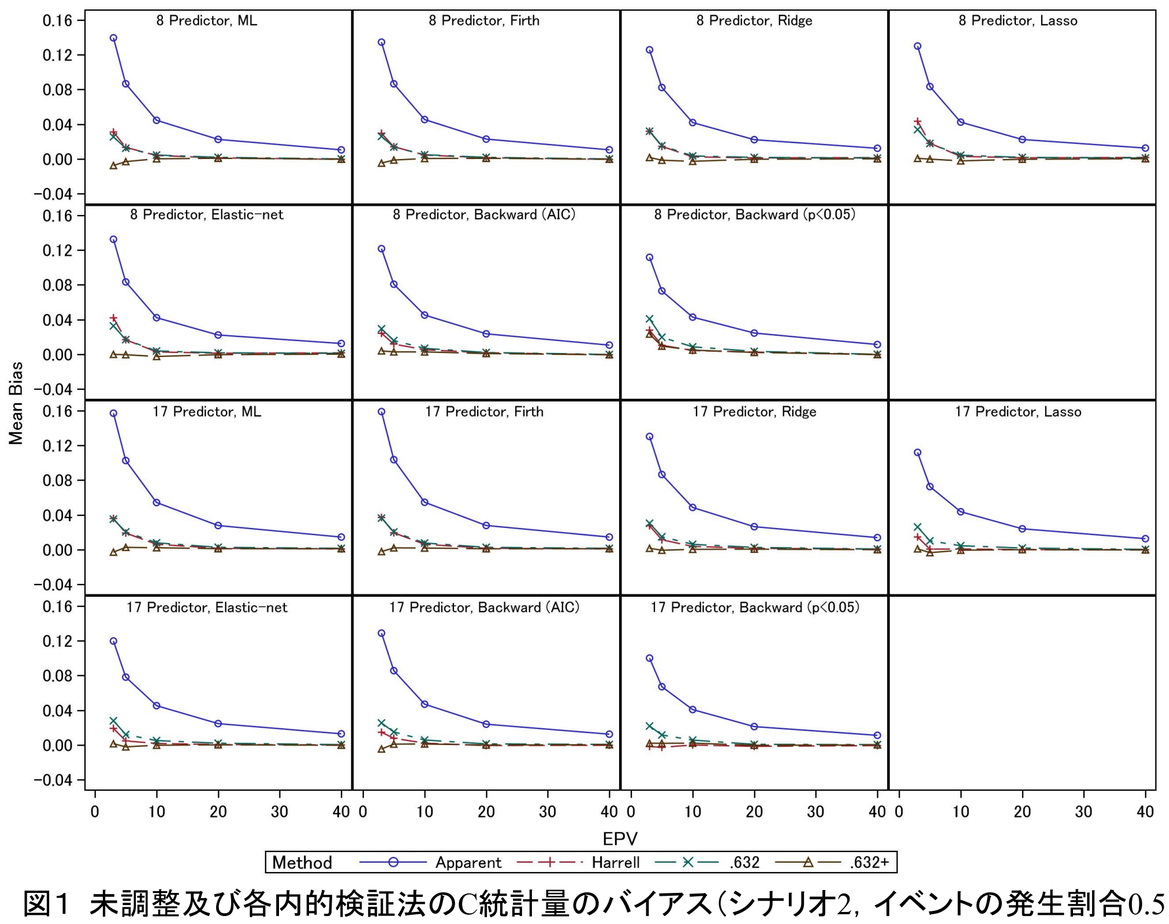

図1　未調整及び各内的検証法のC統計量のバイアス（シナリオ2，イベントの発生割合0.5）

結論

本研究では，近年，臨床研究の実践において普及しつつある正則化法も含めた広範な設定におけるリサンプリング法に基づく内的検証法の性能をシミュレーションで示した．3つのブートストラップ推定量の性能は，一般に同等であったが，小標本では，正則化法が用いられた場合を除いて，.632+推定量の性能が相対的に優れていた．

なお，本研究の成果は，Iba et al.（2021）に掲載されている．

参考文献

Harrell, F. E., Jr., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15, 361-387.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on crossvalidation. *Journal of the American Statistical Association* 78, 316-331.

Efron, B., and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* 92, 548-560.

The GUSTO Investigators. (1993). An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *The New England Journal of Medicine* 329, 673-682.

Iba, K., Shinozaki, T., Maruo, K., and Noma, H. (2021) Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models. *BMC Medical Research Methodology*, 21(1), 9.