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4年

背景

医療において，疾患の診断及び予後の予測は重要な問題の一つである．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．多変量臨床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・較正能力の指標の推定量は，オーバーフィッティングによる過大評価のバイアス(Optimism)を含んでおり，将来予測を行う外部集団での予測能力よりも高くなる傾向がある．Optimismを補正した判別・較正能力を評価するために，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を用いることが推奨されている．このうち，現在では，慣習的に，Harrell et al. (1996) のバイアス補正法が最も広く用いられているが，一方で，Efronの.632法(Efron, 1983)及び.632+法(Efron and Tibshirani, 1997) などの推定量も提案されている．しかしながら，これらの方法の優劣関係を詳細に評価し，いずれの方法が実践において推奨されるかについての評価研究は，これまでに十分に行われていない．本研究では，広範なシミュレーション実験によって，リサンプリング法に基づく内的検証法の比較・評価を行う．特に，臨床予測モデルとして従来のロジスティック回帰(最尤法)，Stepwise法，Firth法，Ridge，Lasso及びElastic-netなど，最新の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，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する．

シミュレーション研究

臨床予測モデルの実践的な状況での内的検証法の評価を行うために，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West regionデータセットに基づいてシミュレーションデータを発生させた．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，臨床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．

予測能力の指標として，予測モデルの総合的な判別性能であるC統計量を用いる．C統計量は，イベントを起こした個体とイベントを起こさなかった個体のペアをランダムに抽出した際，イベントを起こした個体のイベント発生確率の推定値が高くなる確率である．また，これは，イベント発生確率の推定値からイベントの有無を決めるカットオフ値を変化させたときに，感度及び1-特異度をプロットしたROC(receiver operating characteristic)曲線のAUC(area under the curve)に一致する．

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（先行研究で特定された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の設定で検討を行う．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 West regionデータセットに対する最尤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する．切片の真値は，イベントの発生割合に一致するように調整する．

予測変数は，GUSTO-I West regionデータ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を発生させる．各個体のイベント発生確率π_iは，予測変数$\boldsymbol{x}_i$からロジスティックモデル$\pi_i = 1/(1+\exp(-\boldsymbol{\beta}'\boldsymbol{x}_i))$ で求め，イベント変数はベルヌーイ分布Bernoulli(π_i)からの乱数で発生させる．

各設定で，外部集団に対する予測精度を評価するために，50万例のテストデータを発生させる．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る．発生させた学習データを用いて臨床予測モデル(最尤法，Firth法，Ridge，Lasso，Elastic-net，Stepwise法($p<0.05$及びAICに基づく)を構築し，学習データに対するC統計量及びテストデータに対するC統計量を求める．学習データから2000組のブートストラップ標本を発生させ，Harrell法，Efronの.632法及び.632+法によるOptimismを調整したC統計量を求める．テストデータでのC統計量を真値とし，学習データ及び各内的検証法でのC統計量について，バイアスとRMSE(root mean squared error)を評価する．

結果

シナリオ2におけるイベントの発生割合0.5及び0.0625の結果を示した．学習データ及び各内的検証法でのC統計量のバイアスを，図1（イベントの割合0.5）及び図2（イベントの割合0.0625）に示した．全ての推定方法の全ての設定において，学習データのC統計量は相対的に大きな過大評価のバイアスを示した．EPVが10以上では，内的検証法のC統計量はいずれもバイアスがなかった．Harrell法及び.632法は同様の傾向であり，若干の過大評価のバイアスを示した．.632+法は，最尤法及びFirth法において若干の過小評価のバイアスを示すことがあったが，Ridge，Lasso及びElastic-netにおいては，ほぼバイアスを示さなかった．

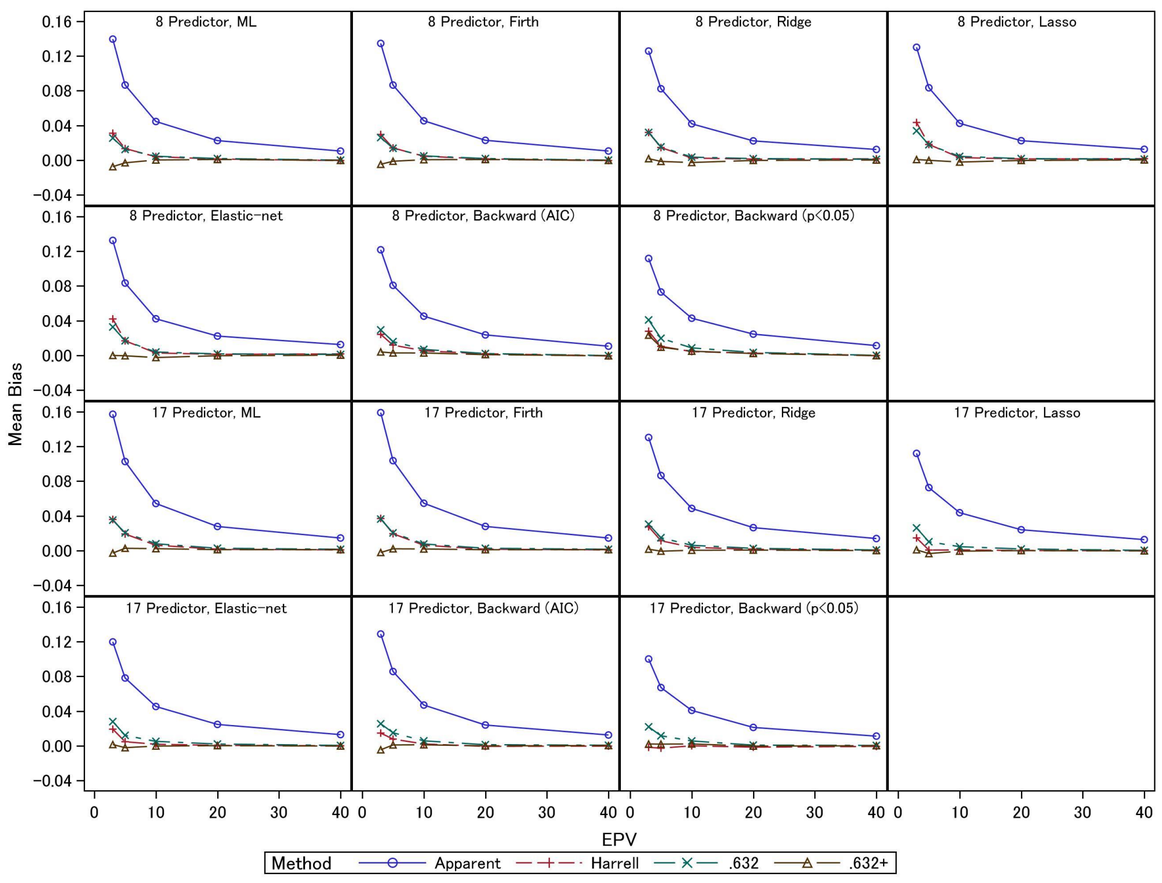

図1 学習データ及び各内的検証法でのC統計量のバイアス（シナリオ 2，イベントの発生割合 0.5）

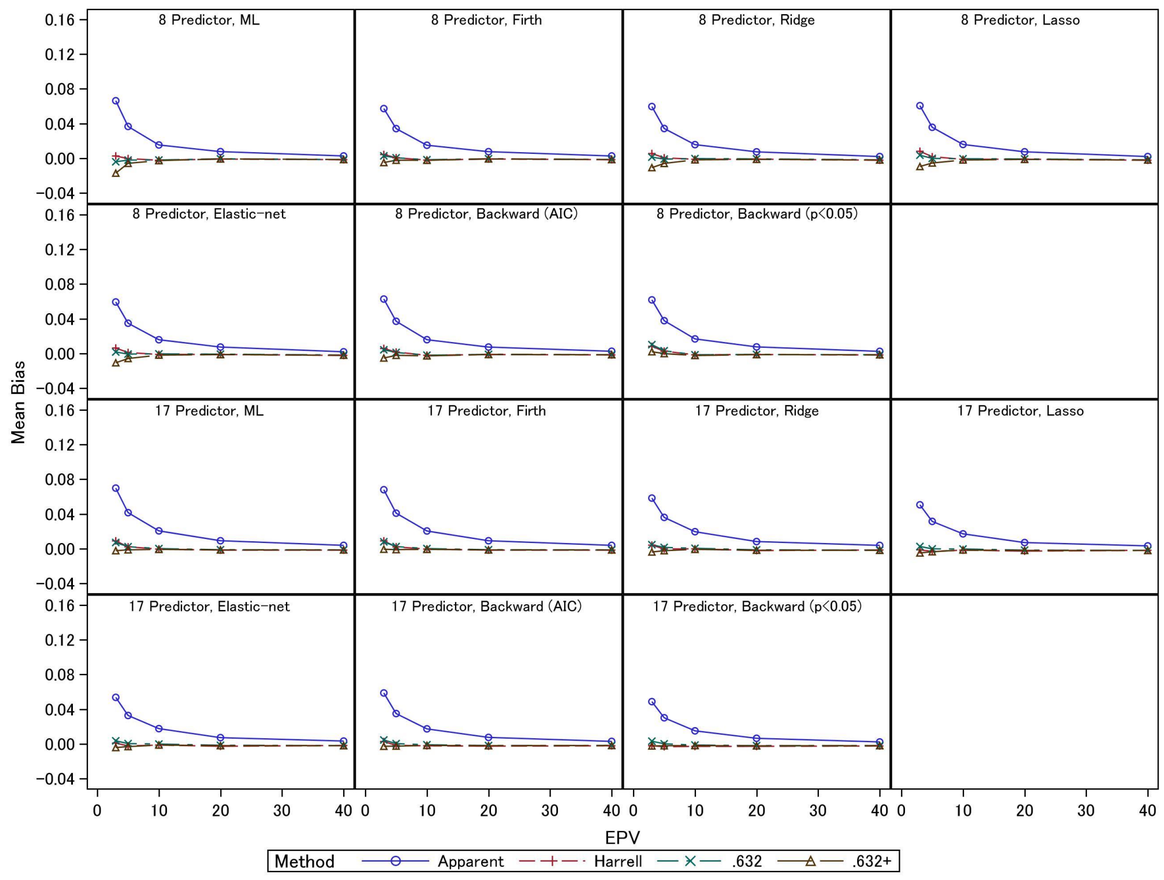

図2 学習データ及び各内的検証法でのC統計量のバイアス（シナリオ 2，イベントの発生割合 0.0625）

結論

ある程度の標本サイズ（EPV 10以上）の状況では，いずれのブートストラップ法に基づく内的検証法も上手く機能した．しかしながら，小標本ではいずれのブートストラップ法に基づく内的検証法にもバイアスがあった．Ridge，Lasso及びElastic-netでは，.632+法のバイアスは相対的に小さかったが，RMSEはHarrell法及び.632法よりも大きかった．それ以外の推定方法では，3つのブートストラップ法に基づく推定量の性能は同様であったが，.632+法が相対的に優れていた．

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.