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

博士（統計科学）

総合研究大学院大学
複合科学研究科
統計科学専攻

令和３（２０２１）年度

概要

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、オプティミズム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の集団に対する実際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。現在のエビデンスおよびガイドラインでは、ブートストラップ法による内的検証法（Harrell 法、Efron の.632 法および.632+法）によって、オプティミズムを補正することが推奨されている。現在、Harrell 法が慣例的に大半の臨床研究で用いられているが、元来、Efron の.632 法および.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、これらの方法の性能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。そのため、広範な実践的条件のもとで、これらの推定量の性能を比較・評価するために、大規模な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、Firth 法、Ridge 回帰、Lasso 回帰および 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。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（Global Utilization of Streptokinase and Tissue plasminogen activator for Occluded coronary artery）試験のデータセットに基づいた設定でシミュレーションデータを生成し、予測変数の数（p）とアウトカムにおけるイベント数（e）の比（e/p）である予測変数あたりのイベント数（events per variables: EPV）、イベ

ント発生割合、候補の予測変数の数、予測変数の回帰係数の真値を変化させることで、広範な実践的条件を考慮した。多変量予測モデルの判別精度の指標として最もよく用いられている C 統計量を性能評価に用いた。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、3 つのブートストラップ法に基づく推定量の性能は、概ね同等であり、いずれにもほとんどバイアスは認められなかった。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+法の性能が相対的に優れていた。したがって、一般的には、現在慣例的に用いられている 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。

また、現在の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おらず、多くの臨床研究において、オプティミスムの補正が行われていない予測精度の指標の信頼区間が報告されている。そのため、オプティミスムを補正した信頼区間の計算方法（位置補正ブートストラップ法および 2 段階ブートストラップ法）を提案した。GUSTO-I 試験のデータセットを基にした設定で行ったシミュレーション実験の結果、従来の未調整の方法は、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いたが、提案法は、どちらの方法も従来の未調整の方法の性能を上回っており、小標本において位置補正ブートストラップ法の被覆確率が名義水準を下回る点を除いては、妥当な信頼区間が得られた。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。

目次

第１章 はじめに

医療の実践において、治療方針などの意思決定を行うために、患者の疾患の診断や予後の予測を行うことは、重要な問題の 1 つである。患者から得られる複数の予測変数に基づいて、疾患の診断およ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[1]。多変量予測モデルは、一般的に、診断や短期的な予後の予測に関する二値アウトカムに対するロジスティック回帰モデルや、長期的な予後の予測に関する生存時間アウトカムに対する Cox 回帰モデルなどの回帰モデルに基づいて構築される。近年では、機械学習の手法などを応用した研究も活発に行われている。多変量予測モデルを開発する際、構築した予測モデルの実践における予測精度を評価することは極めて重要である。しかしながら、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、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[2]。多変量予測モデルの開発および報告に関するガイドラインである TRIPOD（Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis）声明では、予測モデルの構築に用いたデータと同じ母集団から得られる将来のデータに対する予測精度を評価する内的検証法によって、オプティミスムを調整することを推奨している[2, 3]。代表的な内的検証法としては、スプリットサンプル法、クロスバリデーション（CV）法およびブートストラップ法が知られているが、このうち、スプリットサンプル法は、予測モデルの構築に用いるデータを無駄にしており、また、相対的に不正確な推定値を与えるなどの欠点が指摘されている[2, 4]。CV 法は、比較的にバイアスが小さく、精度の高い推定値が得られるが、よく用いられる 10-fold CV 法では、各反復での予測モデルの構築に用いるデータが似通ってい

る（全データの 90%を含んでいる）ことにより、各反復で同じ予測変数が選ばれやすい傾向があるため、ステップワイズ法などの自動変数選択法を用いた場合に生じるモデルの不確実性を適切に反映出来ない可能性がある（これは、各反復で 1 つのデータを除く leave-one-out CV 法ではより明らかである）[4, 5]。したがって、これまでに得られているエビデンスやガイドラインからは、内的検証法として、バイアスが小さく、自動変数選択法によって生じるモデルの不確実性を適切に反映できるブートストラップ法の使用が推奨されている[2, 4]。

ブートストラップ法による代表的なオプティミスムの調整方法として、Harrell のバイアス補正法、Efron の.632 法および.632+法の 3 つの方法が提案されている[1, 6, 7]。現在、比較的にシンプルなアルゴリズムで実行することができる Harrell 法が、慣例的に大半の臨床研究で用いられているが、元来、Efron の.632 法および.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[8]。

これまでに、これらのブートストラップ法に基づく推定量の性能を評価するために、いくつかのシミュレーション研究が実施されている。Steyerberg et al. [4]は、最尤推定による従来のロジスティック回帰モデルを用いた場合におけるこれらの推定量の性能をシミュレーション実験で比較し、一定以上の規模のサンプルサイズのもとで、これらの推定量の性能に大きな違いはなかったと報告している。また、Mondol and Rahman [9]は、ロジスティック回帰モデルに加えて、罰則付き推定の 1 つである Firth 法[10, 11]を含め、イベント発生割合が小さい状況（約 0.1 以下）におけるこれらの推定量の性能評価を行い、.632+法が優れていたと報告している。しかしながら、これらの先行研究では、スプリットサンプル法および CV 法も含めた複数の内的検証法について、複数の予測精度の指標が評価されていたため、ブートストラップ法に基づく推定量の比較

は、Steyerberg et al.では 1 シナリオ、Mondol and Rahman では 3 シナリオの限られた条件のみしか、シミュレーション実験で評価されていなかった。また、近年、臨床研究の実践においても普及しつつある Ridge 回帰[12]、Lasso（least absolute shrinkage and selection operator）回帰[13]および Elastic-net 回帰[14]といった正則化法および現在の実践においても変数選択のために使用されているステップワイズ法[15]を用いた場合の性能評価は行われていなかった。したがって、これらの推定量の性能について、これまでの研究から得られているエビデンスは限定的であり、臨床研究の実践におけるリサンプリング法に基づく内的検証法の使用に関するガイドラインはまだ十分に確立していない。そのため、現在、ほとんどの臨床研究において、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いるが、.632 法や.632+法といった他の方法の実践での有用性についての詳細な分析も、これまでにほとんど行われていない。バイアス補正の正確性・統計的精度は、開発される予測モデルの性能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、医療の実践で行われる診断および予後予測、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。

そのため、本論文では、1 つ目の研究課題として、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、大規模な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 回帰および 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。それによって、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ことを示した。本研究では、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（Global Utilization of Streptokinase and Tissue plasminogen activator for Occluded coronary arteries）試験[16, 17]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。また、本研究では、広範な設定において詳細な分析を行うために、多変量予測モデルの判別精度の指標として最もよく用いられている C 統計量[18]を評価に用いた。他の予測精度の指標への一般化については、考察にて述べる。

また、現在の実践において、内的検証法によるオプティミスムの補正は、主に予測精度の指標の点推定値に対してのみ行われている。多くの臨床研究において、オプティミスムを補正した点推定値が報告されているにも関わらず、信頼区間に関しては、一般的にオプティミスムの補正が行われておらず、C 統計量の場合では、主に従来の DeLong 法[19]による信頼区間のみが示されている。オプティミスムが補正されていない予測精度の指標の推定値には深刻なバイアスがあることから、その信頼区間の実際の被覆確率は、名義水準（一般的に 95%）を大きく下回っていると考えられる。しかしながら、最も標準的な補正法であるブートストラップ法について、信頼区間の補正法はこれまで提案されていない。そのため、本論文の 2 つ目の研究課題として、正則化法なども含む種々のモデル構築法に適用できる 2 つのオプティミスムを補正した推定量に基づく信頼区間の計算方法（位置補正（location-shifted）ブートストラップ法および 2 段階（two-stage）ブートストラップ法）を提案した。シミュレーション実験により、現在用いられている従来の未調整の信頼区間の被覆確率は、顕著に名義水準を下回っているが、3 つのオプティミスム補正法に基づく提案法の信頼区間は、被覆確率を望ましい水準に維持できることを示した。

本論文の構成は、以下のとおりである。第 2 章では、二値アウトカムに対するロジスティック回帰に基づく多変量予測モデルおよび本研究で用いた種々のモデル構築法の概要を説明する。第 3 章では、1 つ目の研究課題であるオプテ

ィミスム補正法の評価に関する研究について述べる。3.1 節では、本研究で予測精度の指標として用いた C 統計量およびブートストラップ法によるバイアス補正法について説明する。3.2 節では、シミュレーション実験の設定の基にした GUSTO-I 試験のデータセットの特性を把握するために、GUSTO-I 試験の実データを用いて、多変量予測モデルの構築およびブートストラップ法に基づく内的検証法による予測精度の評価を行った結果を示す。3.3 節では、最初にシミュレーション実験の設定、データの生成方法および評価方法について説明する。続いて、シミュレーション実験の結果を詳細に分析する。3.4 節では、シミュレーション実験から得られた結果についての考察を行う。第 4 章では、2 つ目の研究課題であるオプティミスムを補正した信頼区間に関する研究について述べる。4.1 節では、提案する位置補正ブートストラップ法および 2 段階ブートストラップ法による信頼区間について説明する。4.2 節では、GUSTO-I 試験の実データに提案法を適用した結果を示す。4.3 節では、提案法の妥当性を確認するために実施したシミュレーション実験について説明する。4.4 節では、シミュレーション実験から得られた結果についての考察を行う。最後に第 5 章で、本論文のまとめを述べる。

第２章　多変量予測モデルおよび種々のモデル構築法

2.1　ロジスティック回帰モデル

最初に、多変量予測モデルおよび本研究で用いる種々のモデル構築法の概要を説明する。二値アウトカム変数に対する回帰モデルに基づく多変量予測モデルとして、一般的にロジスティック回帰モデルが用いられている[2, 5]。患者iの二値アウトカム変数をy_i (= 1: イベント発生 or = 0: 非イベント発生) $(i = 1, 2, \ldots, n)$、p個の予測変数を$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ $(i = 1, 2, \ldots, n)$とする。患者iのイベント発生確率$\pi_i = \Pr(y_i = 1|\boldsymbol{x}_i)$は、ロジスティック回帰モデル

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}$$

によってモデル化される。ここで、$\boldsymbol{\beta} = \left(\beta_0,\ \beta_1, \ldots, \beta_p\right)^T$は、切片を含む回帰係数である。$\boldsymbol{\beta}$に適切な推定値$\widehat{\boldsymbol{\beta}}$をプラグインすることによって得られる患者$i$のイベント発生確率の推定値$\hat{\pi}_i$ $(i = 1, 2, \ldots, n)$は、リスクスコアと呼ばれており、アウトカムを予測する際の基準として用いられる。

$\boldsymbol{\beta}$の最尤推定値$\widehat{\boldsymbol{\beta}}_{ML}$は、対数尤度関数

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log \pi_i + (1 - y_i)\log(1 - \pi_i)\}$$

を最大化することで得られる。この従来の最尤推定値は、標準的な統計解析ソフトウェアで簡単に求めることができ、臨床研究の実践でもよく用いられている[2, 5]。従来の最尤推定量は、大標本のもとでは漸近有効性などの好ましい性質を持つが、小標本またはスパースなデータに対して用いた場合、いくつかの有限標本問題を持つことが知られている。例えば、回帰係数の推定値の絶対値に過大推定のバイアスが生じる可能性がある[20, 21]。また、非常に効果が強い予測変数がある場合、もしくはスパースなデータに用いた場合に、イベント発

生と非イベント発生を完全に分離することができてしまい、イベント発生確率の推定値が 0 または 1 に近づくことによって、回帰係数の推定を行うことができない、もしくは回帰係数の推定値が不安定になることがある[11, 22]。これは（準）完全分離の問題として知られている。これらの問題は、予測変数の数（p）とアウトカムにおけるイベント数（e）の比（e/p）である予測変数あたりのイベント数（events per variables: EPV）が増加することによって見られなくなる。EPV は、多変量予測モデルを構築する際のサンプルサイズの指標として用いられることがあり、慣例的に EPV が 10 以上という基準がよく用いられている[23]。しかしながら、最近の研究において、この基準の妥当性は、様々な条件に依存することが明らかになってきている[24-27]。以下で述べる縮小推定法は、上記の最尤法の問題点を軽減することができる。

アウトカムの予測に寄与しないノイズ変数を多変量予測モデルに含めると、オーバーフィッティングによるオプティミスムが生じる可能性がある。また、実用化の観点からも、少数の予測に寄与する変数のみを含んだ多変量予測モデルの方が有用である。多変量予測モデルに含める変数を選択するために、ステップワイズ法をはじめとした自動変数選択法を用いることができる。ステップワイズ法には、forward 法と backward 法があるが、多変量予測モデルの構築では、一般的に後者の使用が推奨されている[5]。ステップワイズ法の停止基準として、有意水準（慣例的な閾値は $P < 0.05$）、赤池情報量規準（Akaike Information Criterion: AIC）[28]、ベイズ情報量規準（Bayesian Information Criterion: BIC）[29]などが用いられる。AIC と BIC に関して、実践においてどちらの停止基準が良いかについての明確なエビデンスはないが、いくつかの研究において、多変量予測モデルでは AIC が好ましいと報告されていることから[5, 8, 30]、本研究では、AIC のみを採用した。

2.2 Firth 法

Firth 法は、元来、最尤推定量の有限標本バイアスを軽減するために開発された方法である[10, 25]。Firth 法では、罰則項を付与した対数尤度関数

$$l(\boldsymbol{\beta})+\frac{1}{2}\log|I(\boldsymbol{\beta})|$$

によって、回帰係数を推定する。ここで、$I(\boldsymbol{\beta})$は Fisher の情報行列である[11]。Firth 法は、対数尤度関数に罰則項を加えることによって、最尤推定値が無限の値になってしまう（準）完全分離の状況においても、回帰係数を推定することができる[11]。加えて、罰則項は回帰係数の推定値を 0 に向かって縮小するため、小標本またはスパースなデータにおいても、回帰係数の推定値が安定する[25]。また、回帰係数の縮小によってオーバーフィッティングを軽減することができる[24]。

2.3 Ridge 回帰

Ridge 回帰は、予測変数間に強い相関が存在する場合に最尤推定値が不安定になる多重共線性に対処するために開発された方法である[12]。また、Ridge 回帰は、正則化法によって回帰係数の縮小推定を行う代表的な方法の 1 つであり、以下の罰則付き対数尤度関数を最大化することによって、回帰係数の縮小推定を行う。

$$l(\boldsymbol{\beta})-\lambda\sum_{j=1}^{p}\beta_j^2$$

$\lambda\,(>0)$は、縮小の程度を調整するためのチューニングパラメータである。Ridge 回帰の罰則項は回帰係数の平方和であり、回帰係数の推定値が大きくなることに対する罰則によって、回帰係数の推定値が 0 に向かって縮小され、オーバーフィッティングが軽減される[31]。ただし、後述する Lasso 回帰および

Elastic-net 回帰と異なり、Ridge 回帰では、回帰係数を完全に 0 と推定することはできない。なお、Ridge 回帰では、一般的に予測変数はあらかじめ平均 0、分散 1 に標準化される。チューニングパラメータλを選択するために、CV 法など種々の方法が用いられる[24, 31, 32]。

2.4 Lasso 回帰

Lasso 回帰もまた、Ridge 回帰と同様に、正則化法によって回帰係数の縮小推定を行う代表的な方法の 1 つとして、よく用いられている[13]。Lasso 回帰は、回帰係数を推定する際に、回帰係数の絶対値の合計を罰則項として加えた罰則付き対数尤度関数

$$l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \left|\beta_j\right|$$

を用いる。Ridge 回帰と同様に、チューニングパラメータλによって、縮小の程度を調整する。Lasso 回帰は、Ridge 回帰とは異なり、罰則項の特徴により、いくつかの回帰係数を正確に 0 と推定することができ、それによって縮小推定と変数選択を同時に行うことができる。しかしながら、相関の高い予測変数が存在する場合、Lasso 回帰はその内の 1 つの予測変数のみを選択してしまう問題がある[14]。なお、Ridge 回帰と同様に、Lasso 回帰でも予測変数はあらかじめ平均 0、分散 1 に標準化される。チューニングパラメータλは、Ridge 回帰と同様に、CV 法などで選択される。

2.5 Elastic-net 回帰

Elastic-net 回帰は、Lasso 回帰の変数選択の特徴を残したまま、Lasso 回帰の欠点を克服するために提案された正則化法である[14]。Elastic-net 回帰の罰則項

は、Ridge 回帰と Lasso 回帰の罰則項を組み合わせており、罰則付き対数尤度関数は、

$$l(\boldsymbol{\beta}) - \lambda\left\{(1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}\left|\beta_j\right|\right\}$$

で定義される。$\lambda\ (>0)$は、縮小の程度を調整するためのチューニングパラメータであり、$\alpha\ (0 \leq \alpha \leq 1)$は、Ridge 回帰と Lasso 回帰の罰則項の重みを決定するチューニングパラメータである。予測変数間に強い相関がある場合でも、Elastic-net 回帰はそれらの変数を同時にモデルに含めることができる。また、Lasso 回帰と同様に、いくつかの回帰係数を正確に 0 と推定することによって、変数選択と縮小推定を同時に行うことができる[14]。Elastic-net 回帰でも、予測変数はあらかじめ平均 0、分散 1 に標準化される。Elastic-net 回帰では、2 つのチューニングパラメータλおよびαを、2 次元で探索する必要がある。チューニングパラメータを選択する基準としては、Ridge 回帰および Lasso 回帰と同様に、CV 法などが用いられる。

第３章 オプティミスム補正法の評価に関する研究

3.1 C 統計量およびオプティミスム補正法

3.1.1 C 統計量

本節では、本研究で用いた予測精度の指標である C 統計量およびブートストラップ法に基づくオプティミスム補正法について説明する。

本研究では、オプティミスム補正法の性能評価に C 統計量を用いた。C 統計量は、アウトカムの有無を決定するリスクスコアのカットオフ値を変化させた際、横軸に$1-$特異度（擬陽性率）、縦軸に感度（真陽性率）をプロットした ROC（receiver operating characteristic）曲線の AUC（area under the curve）のノンパラメトリックな推定量に対応し[18]、カットオフ値に依らない総合的な判別精度の指標として、多変量予測モデルの判別精度の評価で最もよく用いられている[2]。また、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が高くなる確率$\Pr(\pi_i > \pi_j | y_i = 1, y_j = 0)$に一致する[18]。C 統計量という名称は、$\pi$と$y$の一致度（concordance）に由来している。C 統計量の推定値は

$$\hat{\theta} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} I(\hat{\pi}_i, \hat{\pi}_j)$$

で与えられる[9]。ここで、n_1はイベントを起こした個体数、n_0はイベントを起こさなかった個体数であり、$I(\hat{\pi}_i, \hat{\pi}_j)$は

$$I(\hat{\pi}_i, \hat{\pi}_j) = \begin{cases} 1 & \text{if } \hat{\pi}_i > \hat{\pi}_j \\ 0.5 & \text{if } \hat{\pi}_i = \hat{\pi}_j \\ 0 & \text{if } \hat{\pi}_i < \hat{\pi}_j \end{cases}$$

となる指示関数である。C 統計量は、大きい値ほど判別精度が高いことを意味し、ランダムな予測の場合は 0.5、完璧な予測の場合は 1.0 となる。

3.1.2 Harrell のバイアス補正法

Harrell のバイアス補正法は、従来のブートストラップ法によってオプティミスムを補正する方法であり、現在、慣例的に大半の臨床研究で用いられている [1, 4]。Harrell のバイアス補正法のアルゴリズムを以下に示す。

- 多変量予測モデルの構築に用いたオリジナル標本における未調整の予測精度の指標の推定値を$\hat{\theta}_{app}$とする。
- オリジナル標本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を用いて B 個の予測モデルを構築し、ブートストラップ標本に対する予測精度の指標の推定値 $\hat{\theta}_{1,boot}, \hat{\theta}_{2,boot}, \cdots, \hat{\theta}_{B,boot}$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、オリジナル標本に対する予測精度の指標の推定値$\hat{\theta}_{1,orig}, \hat{\theta}_{2,orig}, \cdots, \hat{\theta}_{B,orig}$を求める。
- オプティミスムのブートストラップ推定値は、上記で得られたブートストラップ標本とオリジナル標本に対する予測精度の指標の推定値の差の平均値である。

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

- 未調整の予測精度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精度の指標の推定値$\hat{\theta}_{app} - \widehat{\Lambda}$を得る。

Harrell のバイアス補正法は、比較的にシンプルなアルゴリズムで計算することができ、一定以上の規模のサンプルサイズ（例えば、EPV ≥ 10）において、

妥当な推定値を与えるという数値的なエビデンスが得られている[4]。そのため、現在、慣例的にほとんどの多変量予測モデルに関する臨床研究において、ブートストラップ法に基づく内的検証法として、Harrell のバイアス補正法が用いられている。しかしながら、オリジナル標本のあるデータが、ブートストラップ標本に含まれる確率は$1-(1-1/n)^n$であることから、nが十分に大きい場合、ブートストラップ標本には、平均的にオリジナル標本の 63.2%のデータが含まれることになる[7]。そのため、ブートストラップ標本とオリジナル標本にデータのオーバーラップが生じており、予測精度の過大推定を引き起こす可能性がある[8]。そのため、以下で述べる標本間のデータのオーバーラップを考慮した代替の推定量が提案されている。

3.1.3 Efron の.632 法

Efron の.632 法[6]は、標本間のデータのオーバーラップを考慮したバイアス補正法である。Efron の.632 法のアルゴリズムは、オリジナル標本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し、各ブートストラップ標本を用いて B 個の予測モデルを構築するところまでは、上述した Harrell のバイアス補正法のアルゴリズムと同じである。.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、この外部標本に対する予測精度の指標の推定値$\hat{\theta}_{1,out},\hat{\theta}_{2,out},\cdots,\hat{\theta}_{B,out}$を求める。データの重複によって実質的なサンプルサイズが小さい（約 0.632・n）ブートストラップ標本を用いて構築された予測モデルは、オリジナル標本の全てのデータを用いて構築された予測モデルよりも予測精度が劣るため、$\hat{\theta}_{out}=\sum_{b=1}^{B}\hat{\theta}_{b,out}/B$をそのままオリジナル標本から構築された予測モデルの予測精度の評価に用いると、過小評価のバイアスが生じ

る。.632 推定量は、この外部標本に対する予測精度の指標の推定値$\hat{\theta}_{out}$の過小評価のバイアスを緩和するために、過大評価のバイアスを含む未調整の予測精度の指標の推定値$\hat{\theta}_{app}$との重み付き平均を取ることによって、

$$\hat{\theta}_{.632} = 0.368 \times \hat{\theta}_{app} + 0.632 \times \hat{\theta}_{out}$$

と定義される。.632 推定量は、モデルのオーバーフィッティングの程度が強くない場合には、概ね不偏推定量となる[6]。なお、ばらつきの小さい$\hat{\theta}_{app}$との重み付き平均を取ることによって、.632 推定量のばらつきは、$\hat{\theta}_{out}$のばらつきよりも小さくなる。.632 推定量の重みは、上述した近似的にブートストラップ標本に含まれるデータの割合に由来している。ブートストラップ標本とブートストラップ標本に含まれなかった外部標本には、データのオーバーラップがないことから、.632 法は CV 法の拡張と考えることができる[4, 9]。しかしながら、予測モデルのオーバーフィッティングの程度が強く、未調整の予測精度の指標の推定値$\hat{\theta}_{app}$の過大推定のバイアスが大きいとき、.632 法はオプティミスムを補正しきれないことが知られている[7]。

3.1.4 Efron の.632+法

Efron and Tibshirani は、予測モデルのオーバーフィッティングの程度を考慮することにより、.632 法の問題点を克服する.632+法[7]を提案した。オーバーフィッティング率Rは、

$$R = \frac{\hat{\theta}_{app} - \hat{\theta}_{out}}{\hat{\theta}_{app} - \gamma}$$

で定義される。γは、無情報モデルの予測精度の指標の推定値であり、オリジナル標本のアウトカム変数をランダムに並び替えたときの予測精度の指標の推定値の平均値で求めることができる。これは、理論的に求めることも可能であり、例えば、C 統計量の場合は、γ ＝0.50 である[4]。オーバーフィッティング

率Rは、オーバーフィッティングがない（$\hat{\theta}_{app} = \hat{\theta}_{out}$）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。.632+推定量[7]は、

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

$$w = \frac{0.632}{1 - 0.368 \times R}$$

で定義される。重みwは、0.632 ($R = 0$)から 1 ($R = 1$)の範囲の値を取る。したがって、.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、外部標本に対する予測精度の指標の推定値$\hat{\theta}_{out}$に近づく。

以下の節の数値計算では、ブートストラップリサンプリングの回数は、$B = 2000$に設定した。また、変数選択およびチューニングパラメータの選択を伴うモデル構築法（ステップワイズ法、Ridge 回帰、Lasso 回帰および Elastic-net 回帰）では、モデル選択の過程で生じる不確実性を適切に考慮するために、モデル構築のすべての手順を、ブートストラップ標本に対するモデル構築においても繰り返した。

3.2 実データの解析

本節では、3.3 節のシミュレーション実験の設定の基にした GUSTO-I 試験のデータセットの特性を把握するために、GUSTO-I 試験の実データを用いて、多変量予測モデルの構築およびブートストラップ法に基づく内的検証法による予測精度の評価を行った結果を示す。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り[16]、これまでに、Steyerberg et al.の先行研究をはじめとして、多変量予測モデルの性能を評価した複数の研究でも用いられている[4, 15, 32]。また、GUSTO-I 試験は、多変量予測モデルに関する代表的な教科書[5]でも、教科書全

体を通して主要な事例として用いられており、多変量予測モデルの分野における代表的な 2 値データの事例である。本研究では、GUSTO-I 試験の一部分である Western データセット[5]を用いた。GUSTO-I 試験 Western データセットには、2188 例の患者のデータが含まれている。二値アウトカム変数は、心筋梗塞の発症後 30 日の死亡の有無であり、17 個の予測変数のデータが得られている。アウトカム変数および 17 個の予測変数の要約を表 3.2-1 に示した。17 個の予測変数のうち、2 変数（身長および体重）は連続変数、1 変数（喫煙歴）は順序変数であり、残りの 14 変数は二値変数である。なお、年齢は連続変数を 65 歳で二値化している。また、喫煙歴は、3 カテゴリー（喫煙者、前喫煙者、非喫煙者）から 2 つのダミー変数（喫煙者 vs. 非喫煙者および前喫煙者 vs. 非喫煙者）を作成し、これらのダミー変数を解析に用いた。

本研究では、 GUSTO-I 試験を用いた複数の先行研究で採用されていた 8 変数モデル（年齢、性別、糖尿病、低血圧、頻脈、高リスク、ショックおよび胸痛緩和なし）[4, 33]、および表 3.2-1 の全ての変数を用いた 17 変数モデルを考慮した。8 変数モデルおよび 17 変数モデルの EPV は、それぞれ 16.9 および 7.5 である。これらの 2 つのモデルに対して、第 2 章で説明したモデル構築法（最尤法、Firth 法、Ridge 回帰、Lasso 回帰、Elastic-net 回帰および backward のステップワイズ法）を用いて、多変量予測モデルを構築した（ステップワイズ法の停止基準には、AIC および $P < 0.05$ を用いた）。解析には、すべて R のバージョン 3.5.1[34]を用いた。最尤法による従来のロジスティック回帰の当てはめには、`glm` 関数を用いた。Firth 法は、`logistf` パッケージ[35]で実行した。Ridge 回帰、Lasso 回帰および Elastic-net 回帰は、`glmnet` パッケージ[36]を用い、チューニングパラメータの選択は、逸脱度（deviance）を評価指標とした 10-fold CV 法で決定した。ステップワイズ法は、`stats` および `logistf` パッケージ[35]を用いた。そして、構築した多変量予測モデルについて、未調整お

よび 3.1 節で説明したバイアス補正法によってオプティミスムを調整した C 統計量を算出した。8 変数モデルの結果を表 3.2-2 に、17 変数モデルの結果を表 3.2-3 に示した。

表 3.2-1 GUSTO-I 試験 Western データセットの要約

N	2188
アウトカム変数	
心筋梗塞の発症後 30 日の死亡	6.2%
予測変数	
年齢>65 歳	38.4%
性別，女性	24.9%
糖尿病	14.3%
低血圧（収縮期血圧<100mmHg）	9.6%
頻脈（脈拍数>80bpm）	33.4%
高リスク（前壁梗塞/心筋梗塞の既往）	48.7%
ショック（Killip 分類 III/IV）	1.5%
胸痛緩和までの時間>1 時間	60.9%
心筋梗塞の既往	17.1%
身長（cm）	172.1 ± 10.1
体重（kg）	82.9 ± 17.7
高血圧の既往	40.4%
喫煙歴，前喫煙者	30.8%
喫煙歴，喫煙者	27.9%
高コレステロール血症	40.5%
狭心症の既往	34.1%
心筋梗塞の家族歴	47.6%
ST 上昇>4 誘導	35.6%

8 変数モデルでは、Lasso 回帰および Elastic-net 回帰は、全 8 変数を選択した。また、2 つのステップワイズ法は、同じ 6 変数（糖尿病および胸痛緩和なし以外）を選択した。17 変数モデルでは、Lasso 回帰および Elastic-net 回帰は、同じ 12 変数（糖尿病、身長、高血圧の既往、前喫煙者、高コレステロール血症および心筋梗塞の家族歴以外）を選択した。ステップワイズ法は、AIC

では 9 変数、P < 0.05 では 7 変数を選択した（表 3.2-3 参照）。8 変数モデルおよび 17 変数モデルの両方で、ステップワイズ法の選択した予測変数の数は、Lasso 回帰および Elastic-net 回帰の選択した予測変数の数よりも少なかった。Firth 法および正則化法（Ridge 回帰、Lasso 回帰および Elastic-net 回帰）では、縮小推定により、最尤法と比較して、回帰係数の推定値の絶対値が小さくなる傾向が認められた。

17 変数モデルの未調整の C 統計量は、8 変数モデルの未調整の C 統計量よりも全体的に大きかった。8 変数モデルでは、全てのモデル構築法の未調整の C 統計量は同程度（約 0.82）であったが、17 変数モデルの未調整の C 統計量は 0.82 から 0.83 の範囲であった。全てのモデル構築法において、オプティミスムを調整した C 統計量（約 0.81）は、未調整の C 統計量よりも小さく、未調整の C 統計量に過大評価のバイアスがあることを示唆した。3 つのブートストラップ法によるオプティミスムを調整した C 統計量は、いずれのモデル構築法でも差が認められなかった。8 変数モデルおよび 17 変数モデルのオプティミスムを調整した C 統計量は同程度であり、ノイズ変数を含んでいる 17 変数モデルの方が、オプティミスムが大きいことが示唆された。

表 3.2-2 8 変数モデルの回帰係数の推定値および C 統計量

	最尤法	Firth 法	Ridge 回帰	Lasso 回帰	Elastic-net 回帰	ステップワイズ法 (AIC)	ステップワイズ法 (P<0.05)
回帰係数:							
切片	-5.092	-5.034	-4.787	-4.933	-4.886	-4.927	-4.927
年齢>65 歳	1.637	1.616	1.424	1.578	1.535	1.631	1.631
女性	0.622	0.620	0.592	0.586	0.586	0.624	0.624
糖尿病	0.069	0.083	0.078	0.024	0.035	.	.
低血圧	1.218	1.215	1.102	1.164	1.145	1.252	1.252
頻脈	0.650	0.645	0.574	0.608	0.597	0.661	0.661
高リスク	0.847	0.835	0.748	0.796	0.781	0.855	0.855
ショック	2.395	2.362	2.339	2.362	2.354	2.424	2.424
胸痛緩和なし	0.263	0.255	0.237	0.219	0.221	.	.
C 統計量:							
未調整	0.819	0.819	0.819	0.819	0.819	0.820	0.820
Harrell	0.810	0.810	0.812	0.810	0.810	0.811	0.810
.632	0.811	0.811	0.812	0.811	0.810	0.811	0.809
.632+	0.810	0.811	0.812	0.811	0.810	0.811	0.809

表 3.2-3 17 変数モデルの回帰係数の推定値および C 統計量

	最尤法	Firth 法	Ridge 回帰	Lasso 回帰	Elastic-net 回帰	ステップワイズ法 (AIC)	ステップワイズ法 (P<0.05)
回帰係数:							
切片	-5.090	-4.983	-3.434	-3.494	-3.486	-3.494	-2.853
年齢>65 歳	1.429	1.399	1.161	1.336	1.324	1.495	1.532
女性	0.490	0.487	0.380	0.288	0.289	0.368	.
糖尿病	0.153	0.164	0.131	.	.	.	.
低血圧	1.192	1.178	1.037	1.036	1.030	1.230	1.227
頻脈	0.653	0.643	0.533	0.530	0.526	0.669	0.717
高リスク	0.403	0.397	0.390	0.372	0.372	0.414	2.748
ショック	2.685	2.608	2.508	2.460	2.455	2.662	0.779
胸痛緩和なし	0.233	0.223	0.200	0.107	0.107	.	.
心筋梗塞の既往	0.505	0.495	0.437	0.378	0.376	0.586	.
身長	0.008	0.008	-0.001	.	.	.	.
体重	-0.019	-0.018	-0.015	-0.014	-0.014	-0.018	-0.023
高血圧の既往	-0.165	-0.159	-0.123	.	.	.	.
前喫煙者	0.174	0.169	0.147	.	.	.	.
喫煙者	0.247	0.241	0.231	0.057	0.059	.	.
高コレステロール血症	-0.064	-0.060	-0.064	.	.	.	.
狭心症の既往	0.246	0.243	0.235	0.151	0.151	.	.
心筋梗塞の家族歴	-0.015	-0.014	-0.036	.	.	.	.
ST 上昇>4 誘導	0.583	0.571	0.479	0.434	0.430	0.601	0.752
C 統計量:							
未調整	0.832	0.832	0.831	0.831	0.831	0.829	0.824
Harrell	0.811	0.811	0.812	0.812	0.812	0.810	0.806
.632	0.811	0.811	0.813	0.813	0.813	0.809	0.808
.632+	0.810	0.810	0.812	0.812	0.812	0.809	0.808

3.3 シミュレーション実験

3.3.1 シミュレーション実験の方法

3.3.1.1 シミュレーション実験の設定

ブートストラップ法に基づく内的検証法の性能を評価するために、3.2 節で解析した GUSTO-I 試験のデータセットに基づいた広範な設定のもとで、シミュレーションデータを生成した。予測精度に影響する可能性がある要因として、EPV（3、5、10、20 および 40）、イベント発生割合（0.5、0.25、0.125 および 0.0625）、候補の予測変数の数（先行研究で用いられた 8 変数および全 17 変数）および予測変数の回帰係数（GUSTO-I 試験 Western データセットに対する最尤推定値（係数タイプ 1）および Elastic-net 回帰の縮小推定値（係数タイプ 2））を考慮した。これらの要因を組み合わせた合計 80 のシナリオで検討を行った。EPV およびイベント発生割合の設定は、多変量予測モデルに関する先行のシミュレーション研究[4, 24]の設定を参考にした。Steyerberg et al. [4]の結果から、EPV ≥ 40 ではオプティミスムがほとんどないと予想されたため、EPV の上限は 40 に設定した。また、本研究では正則化法のチューニングパラメータを 10-fold CV 法で決定しているが、後で述べるように、EPV = 3 において Lasso 回帰が切片のみの予測モデルとなる割合が 20%を超えるシナリオが認められたことから、本研究で用いたモデル構築法が概ね実行可能である EPV = 3 を下限に設定した。予測変数の回帰係数（切片β_0を除く）の設定は、係数タイプ 1 では、全ての予測変数にイベント発生のリスクに関する一定の効果があると仮定し、係数タイプ 2 では、予測変数の効果が相対的に小さく、いくつかの予測変数がイベント発生のリスクに寄与しないと仮定した。切片β_0の真値の設定によって、イベント発生割合を調整した。予測モデルの構築に用いるデータセットのサンプルサイズ n は、（候補の予測変数の数 × EPV / イベント発生割合）で算出した。

3.3.1.2 シミュレーションデータの生成方法

予測変数のデータは、GUSTO-I 試験 Western データセットから推定したパラメータに基づいて、乱数で生成した。3 つの連続変数（年齢、身長および体重）は、GUSTO-I 試験 Western データセットと同じ平均ベクトルおよび分散共分散行列の多変量正規分布からの乱数で生成した。その後、3.2 節の実データの解析と同様に、年齢は 65 歳で二値化した。順序変数の喫煙歴は、各カテゴリの比率が GUSTO-I 試験 Western データセットでの各カテゴリの割合と同じ多項分布からの乱数で生成した。その後、3.2 節の実データの解析と同様に、2 つのダミー変数に変換した。残りの二値変数は、GUSTO-I 試験 Western データセットと同じ周辺確率および相関係数行列の多変量二項分布[37]からの乱数で生成した。多変量二項分布からの相関のある二値変数の乱数生成には、R の `mipfp` パッケージ[38]を用いた。イベント発生確率π_i $(i = 1, 2, \ldots, n)$は、予測変数$\boldsymbol{x}_i$からロジスティック回帰モデル$\pi_i = 1/(1 + \exp(-\boldsymbol{\beta}^T \boldsymbol{x}_i))$に基づいて決定した。アウトカム変数$y_i$は、イベント発生確率$\pi_i$のベルヌーイ分布からの乱数で生成した。

3.3.1.3 シミュレーション実験の評価方法

構築した多変量予測モデルの外部標本に対する実際の予測精度を評価するために、500,000 例の検証データセットを独立に発生させた。この外部標本に対する C 統計量（以下、外部の C 統計量）が estimand である。シミュレーション実験の反復回数は、全てのシナリオで 2000 回に設定した。シミュレーション実験では、各反復で予測モデルの構築に用いるデータセットを生成し、7 つのモデル構築法（最尤法、Firth 法、Ridge 回帰、Lasso 回帰、Elastic-net 回帰および backward ステップワイズ法（AIC および $P < 0.05$））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 統計量お

および 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 法および.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、未調整および各内的検証法の C 統計量のバイアスおよび RMSE（root mean squared error）を評価した。

上記のシミュレーション実験では、3 つの連続変数（年齢、身長および体重）を多変量正規分布からの乱数で生成した。臨床研究から得られるデータの中には、左右対称ではない分布に従う予測変数もあると考えられるため、連続変数の分布の歪みがシミュレーション実験の結果に影響するかどうかを検討するために、GUSTO-I 試験 Western データセットから推定したパラメータの多変量歪正規分布（multivariate skew normal distribution）[39]からの乱数で、3 つの連続変数を生成した。多変量歪正規分布は、分布の歪みを考慮できるように従来の多変量正規分布を一般化した分布であり、位置および尺度パラメータに加えて、各変数の分布の歪みの程度を表す歪度パラメータを含んでいる。歪度パラメータの値が 0 のとき、正規分布と同様に分布の歪みがなく、歪度パラメータの値が正（負）のとき、分布が正（負）の方向に歪んでいる。多変量歪正規分布からの乱数生成には、R の sn パッケージ[40]を用いた。多変量歪正規分布に基づくシミュレーション実験では、感度分析として、最尤法で構築した予測モデルのみ評価した。

3.3.2 シミュレーション実験の結果

以下で述べるように、シミュレーション実験の結果の全体的な傾向は大きく異ならないことから、以下の各セクションでは、回帰係数の設定が係数タイプ 2 で、イベント発生割合 0.5 および 0.0625 の場合を、代表的な結果として詳細に示す。それ以外の結果については、概略を示す。

小標本のシナリオ（EPV が小さく、イベント発生割合が大きい）において、変数選択を伴うモデル構築法（Lasso 回帰、Elastic-net 回帰およびステップワイズ法）は、全ての予測変数を除外してしまい、切片のみの無意味な予測モデルとなってしまうことがあった。実践では、切片のみの予測モデルが最終のモデルとして採用されることはないと考えられるため、切片のみの予測モデルとなったケースは性能評価から除いた。切片のみの予測モデルの発生頻度を表 3.3.2-1（Lasso 回帰および Elastic-net 回帰）および表 3.3.2-2（ステップワイズ法）に示した。切片のみの予測モデルは、EPV = 20 以上では発生せず、EPV = 10 では稀に発生した。切片のみの予測モデルの発生頻度は、EPV = 3 かつイベント発生割合 = 0.5 のもとでの 8 変数モデルで高かった。切片のみの予測モデルが高頻度で発生した原因として、このシナリオは全てのシナリオの中で最もサンプルサイズが小さく（$n = 44$）、変数選択法が上手く機能しなかったことが考えられた。そのため、小標本では、変数選択法を利用できない場合があることが示唆された。そのような場合でも、Firth 法および Ridge 回帰による縮小推定は利用可能である。

表 3.3.2-1 切片のみの予測モデルの割合（%）（Lasso 回帰および Elastic-net 回帰）

		Lasso 回帰				Elastic-net 回帰			
	イベント	8 変数モデル		17 変数モデル		8 変数モデル		17 変数モデル	
EPV	発生割合	C1	C2	C1	C2	C1	C2	C1	C2
3	0.5	17.95	20.50	2.45	5.60	9.80	12.60	1.00	2.05
3	0.25	5.35	7.60	0.00	0.40	2.90	4.20	0.00	0.10
3	0.125	1.50	3.00	0.00	0.00	0.50	1.25	0.00	0.00
3	0.0625	1.05	1.30	0.00	0.10	0.40	0.35	0.00	0.00
5	0.5	3.80	5.70	0.15	0.50	1.90	3.10	0.00	0.00
5	0.25	0.85	1.35	0.00	0.00	0.40	0.90	0.00	0.00
5	0.125	0.25	0.55	0.00	0.00	0.15	0.25	0.00	0.00
5	0.0625	0.00	0.10	0.00	0.00	0.00	0.05	0.00	0.00
10	0.5	0.10	0.15	0.00	0.00	0.00	0.00	0.00	0.00
10	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.125	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.0625	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

C1: 係数タイプ 1, C2: 係数タイプ 2

表 3.3.2-2 切片のみの予測モデルの割合（%）（ステップワイズ法）

		ステップワイズ法（AIC）				ステップワイズ法（P<0.05）			
	イベント	8 変数モデル		17 変数モデル		8 変数モデル		17 変数モデル	
EPV	発生割合	C1	C2	C1	C2	C1	C2	C1	C2
3	0.5	1.15	1.85	0.00	0.05	11.20	14.10	0.15	0.65
3	0.25	0.30	0.55	0.00	0.00	2.50	4.20	0.00	0.05
3	0.125	0.00	0.05	0.00	0.00	0.40	1.10	0.00	0.00
3	0.0625	0.05	0.15	0.00	0.00	0.25	0.65	0.00	0.00
5	0.5	0.15	0.30	0.00	0.00	2.25	3.80	0.00	0.00
5	0.25	0.00	0.00	0.00	0.00	0.20	0.45	0.00	0.00
5	0.125	0.00	0.00	0.00	0.00	0.05	0.10	0.00	0.00
5	0.0625	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.5	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
10	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.125	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.0625	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

C1: 係数タイプ 1, C2: 係数タイプ 2

3.3.2.1 未調整、外部および各内的検証法の C 統計量の結果

係数タイプ 2 かつイベント発生割合 0.5 および 0.0625 における未調整、外部および各内的検証法の C 統計量の平均値を図 3.3.2.1-1 および図 3.3.2.1-2 に示した。なお、すべてのシナリオを通しての各 C 統計量の平均値のモンテカルロ標準誤差の最大値は 0.0023 であった。

イベント発生割合 0.5 では、検証データセットに対する外部の C 統計量は、EPV が 3〜5 では 0.65-0.70 程度であり、大きな EPV では 0.72 付近であった。この結果は、8 変数モデルおよび 17 変数モデルで同様であった。EPV = 3 のもとでの 17 変数モデルでは、Ridge 回帰、Lasso 回帰および Elastic-net 回帰は、他のモデル構築法と比較して、大きな外部の C 統計量（0.67）を示した。最尤法、Firth 法およびステップワイズ法（AIC）の外部の C 統計量は、同程度（0.66）であった。ステップワイズ法（P < 0.05）は、最も小さい外部の C 統計量（0.65）を示した。EPV = 3 のもとでの 8 変数モデルでは、Ridge 回帰、Elastic-net 回帰および Firth 法の外部の C 統計量は、同程度（0.68）であった。しかしながら、Lasso 回帰の外部の C 統計量（0.67）は、最尤法の外部の C 統計量に近かった。両方のステップワイズ法は、他のモデル構築法と比較して、小さな外部の C 統計量（0.64-0.66）を示した。

一般的に、正則化法は、他のモデル構築法よりも良好な実際の予測精度を示した。特に、ノイズ変数を含んでいる 17 変数モデルにおいて、Ridge 回帰、Lasso 回帰および Elastic-net 回帰は、Firth 法と比較して、実際の予測精度が高かった。しかしながら、8 変数モデルでは、Lasso 回帰の実際の予測精度は、Firth 法よりも若干低かった。この原因として、8 変数モデルのシナリオでのサンプルサイズが、17 変数モデルのシナリオでのサンプルサイズよりも小さかったことが考えられた。8 変数モデルでは、Firth 法は最尤法よりも良好な実際の予測精度を示したが、17 変数モデルでは、Firth 法と最尤法の実際の予測精度

は同程度であった。また、先行研究[24]でも見られたように、ステップワイズ法の実際の予測精度は、一般的に他のモデル構築法よりも悪かった。モデル構築法間の実際の予測精度の差は、EPV が大きくなるにつれて小さくなった。

同様の傾向がイベント発生割合 0.0625 でも認められたが、外部の C 統計量は、全体的にイベント発生割合 0.5 よりも高かった（EPV が 3〜5 で 0.75 付近）。これは、EPV が同じ場合、イベント発生割合が小さいシナリオのサンプルサイズが大きいことから、サンプルサイズの違いに起因していると考えられた。モデル構築法の比較に関しては、イベント発生割合 0.5 の場合と同様の傾向であり、正則化法の実際の予測精度が他のモデル構築法よりも高い傾向があったのに対し、ステップワイズ法の実際の予測精度は、他のモデル構築法よりも低い傾向があった。

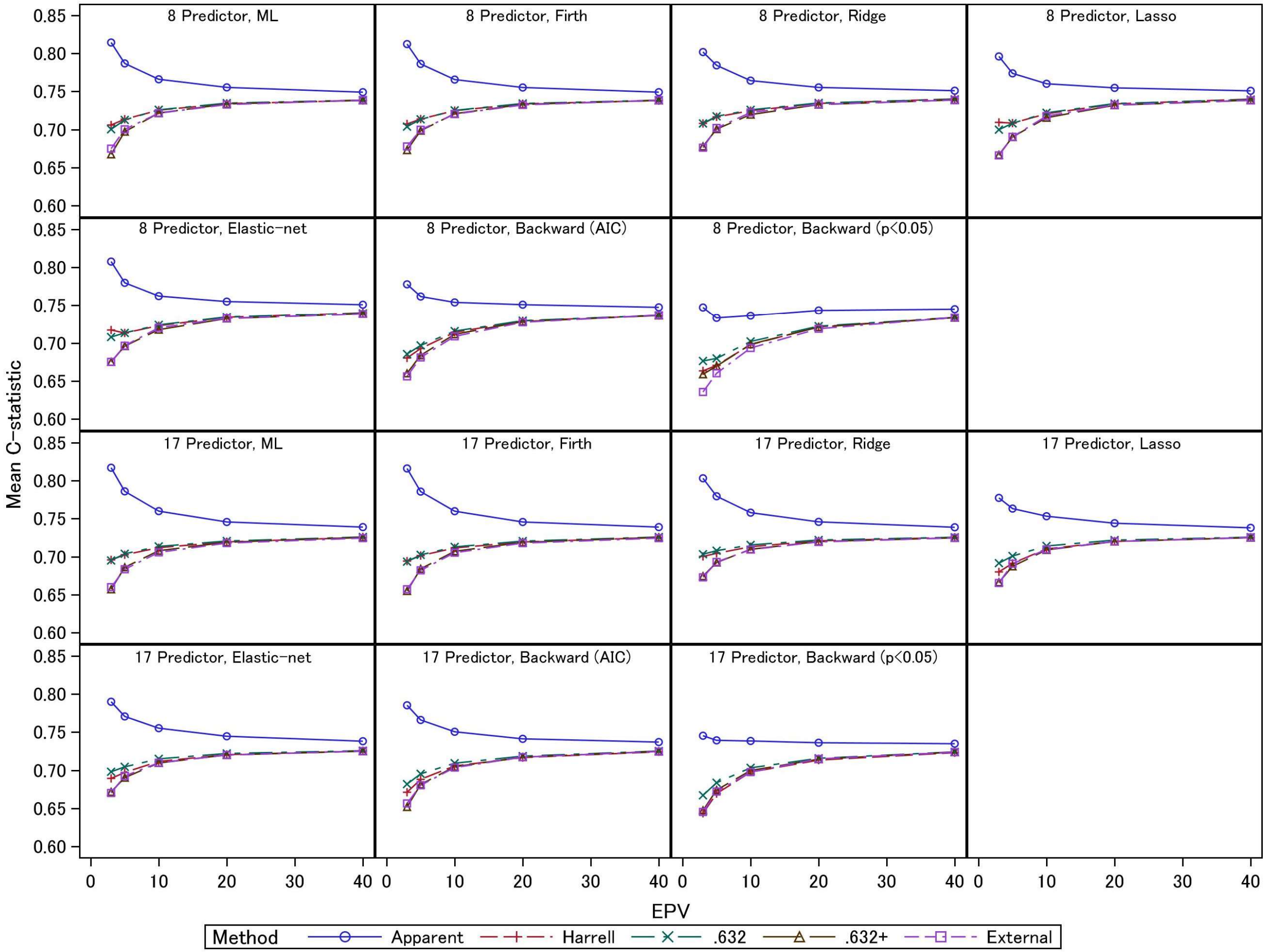

図 3.3.2.1-1 未調整、外部および各内的検証法の C 統計量（係数タイプ 2、イベント発生割合 0.5）

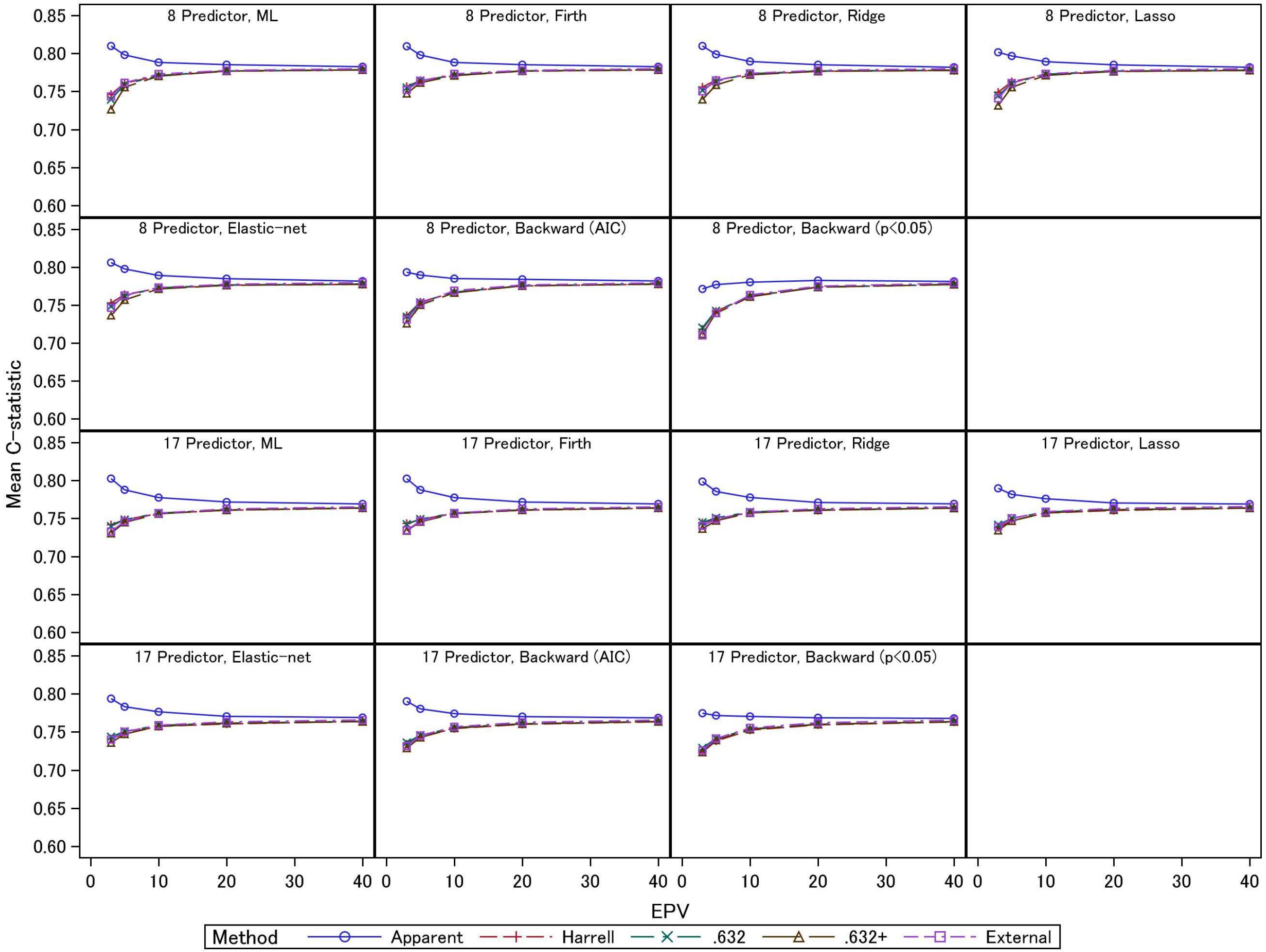

図 3.3.2.1-2 未調整、外部および各内的検証法の C 統計量（係数タイプ 2、イベント発生割合 0.0625）

係数タイプ 1 のイベント発生割合 0.5、0.25、0.125、0.0625 および係数タイプ 2 のイベント発生割合 0.25、0.125 における未調整、外部および各内的検証法の C 統計量の平均値を図 3.3.2.1-3～図 3.3.2.1-8 に示した。

係数タイプ 2 のイベント発生割合 0.25 および 0.125 における外部の C 統計量は、イベント発生割合 0.5 よりも高く、イベント発生割合が小さくなるにつれ、外部の C 統計量が大きくなる傾向が認められた。モデル構築法の比較に関しては、イベント発生割合 0.5 および 0.0625 の場合と同様の傾向であり、正則化法は、一般的に他のモデル構築法よりも良好な実際の予測精度を示した。最尤法と比較して、Firth 法の実際の予測精度が若干高い傾向が認められた。ステップワイズ法は、特に $P < 0.05$ の基準を用いた場合に、他のモデル構築法よりも実際の予測精度が低かった。

回帰係数の真値に GUSTO-I 試験 Western データセットの最尤推定値を設定した係数タイプ 1 のシナリオでは、Elastic-net 回帰の縮小推定値を設定した係数タイプ 2 のシナリオと比較して、外部の C 統計量は全体的に大きかった。これは、係数タイプ 1 の回帰係数の真値が、係数タイプ 2 の回帰係数の真値よりも大きく、各予測変数のイベント発生のリスクへの寄与が大きくなったことを反映していた。また、各モデル構築法間の実際の予測精度の比較は、係数タイプ 2 のシナリオで認められた傾向と大きく異ならなかったが、変数選択を伴うモデル構築法の実際の予測精度は、相対的に若干低下した。この原因として、各予測変数のイベント発生のリスクへの寄与が増加したことにより、変数選択を伴うモデル構築法と変数選択を伴わないモデル構築法間の差が小さくなったことが考えられた。

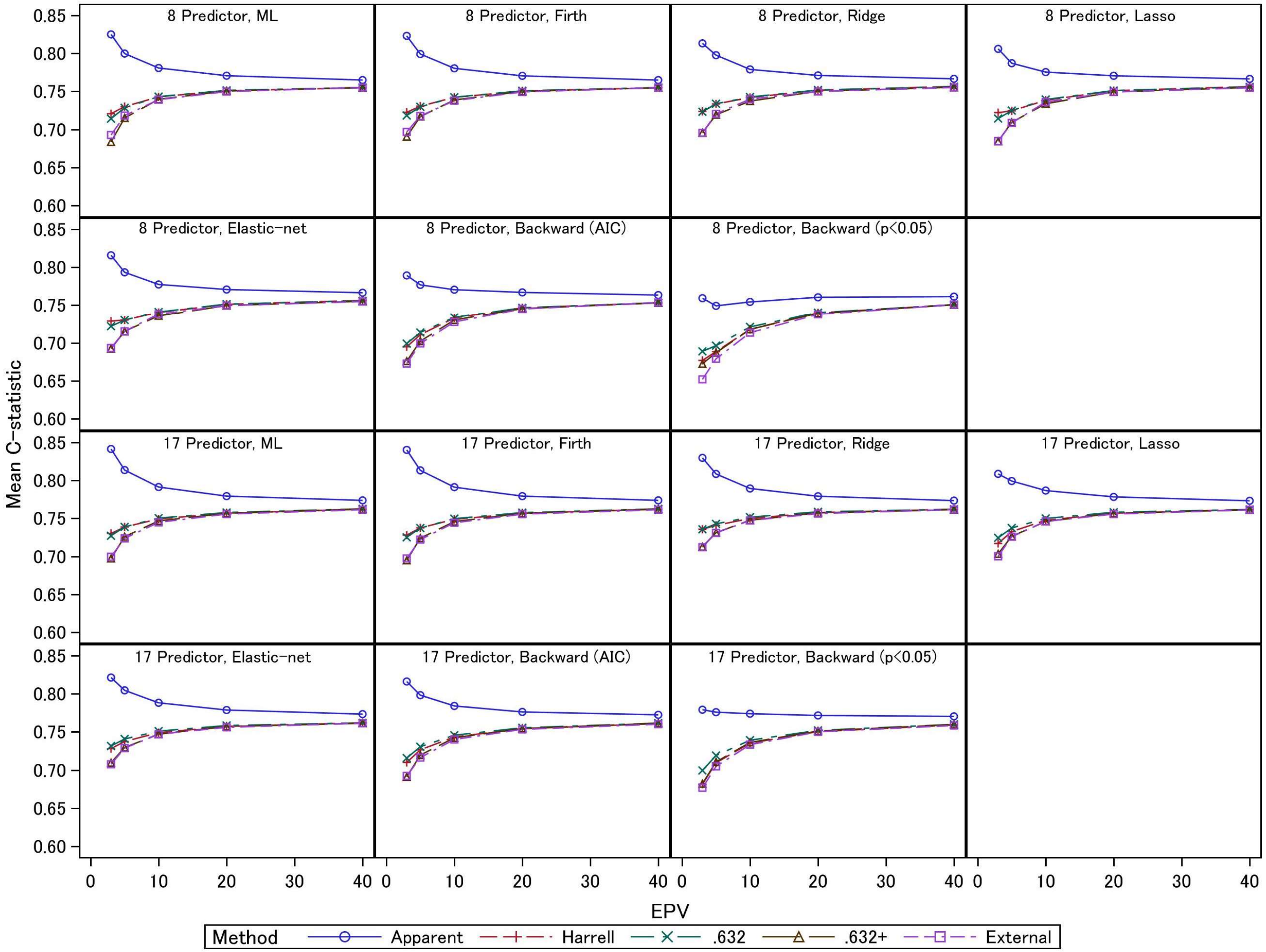

図 3.3.2.1-3 未調整、外部および各内的検証法の C 統計量（係数タイプ 1、イベント発生割合 0.5）

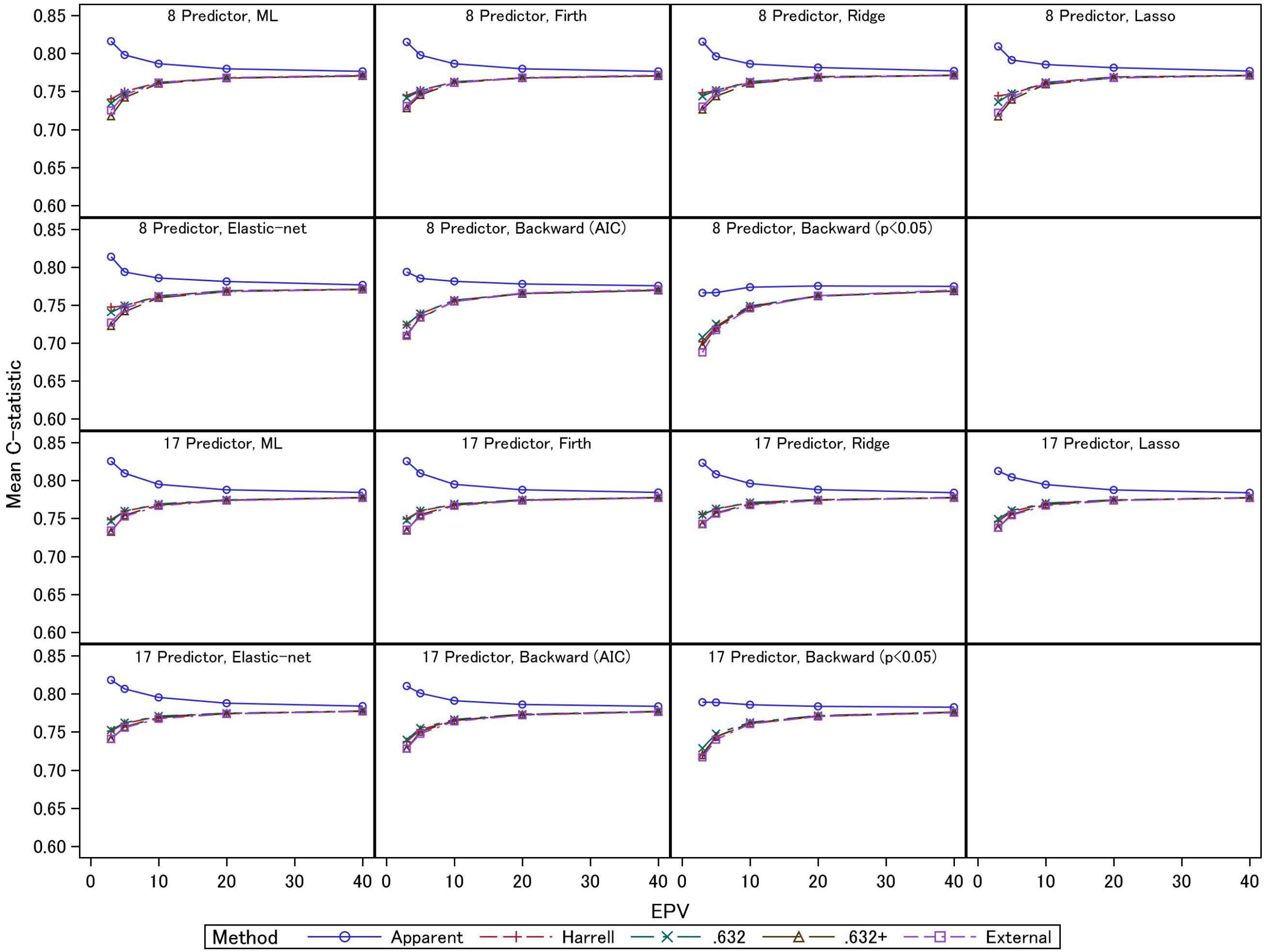

図 3.3.2.1-4 未調整、外部および各内的検証法の C 統計量（係数タイプ 1、イベント発生割合 0.25）

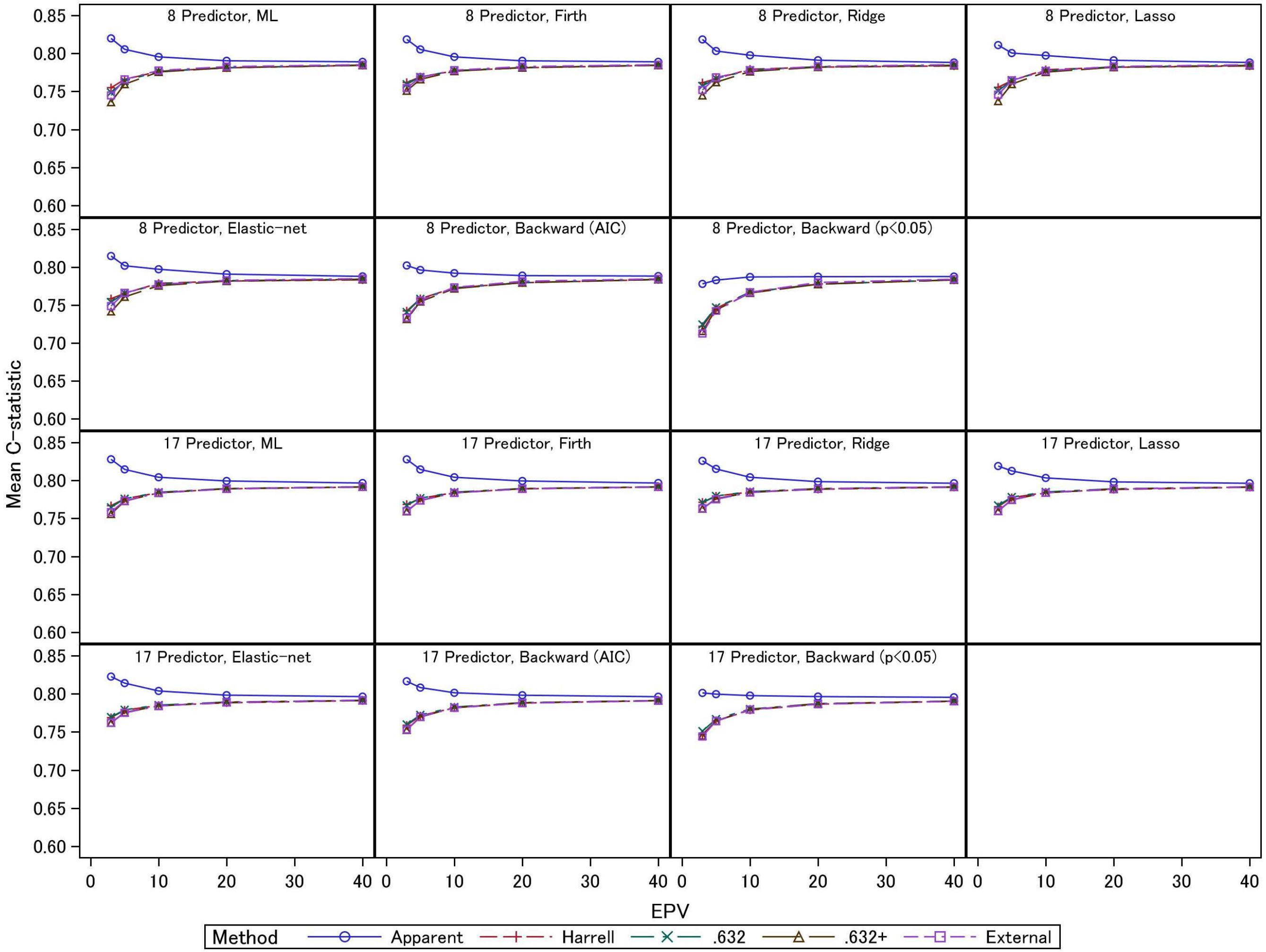

図 3.3.2.1-5 未調整、外部および各内的検証法の C 統計量（係数タイプ 1、イベント発生割合 0.125）

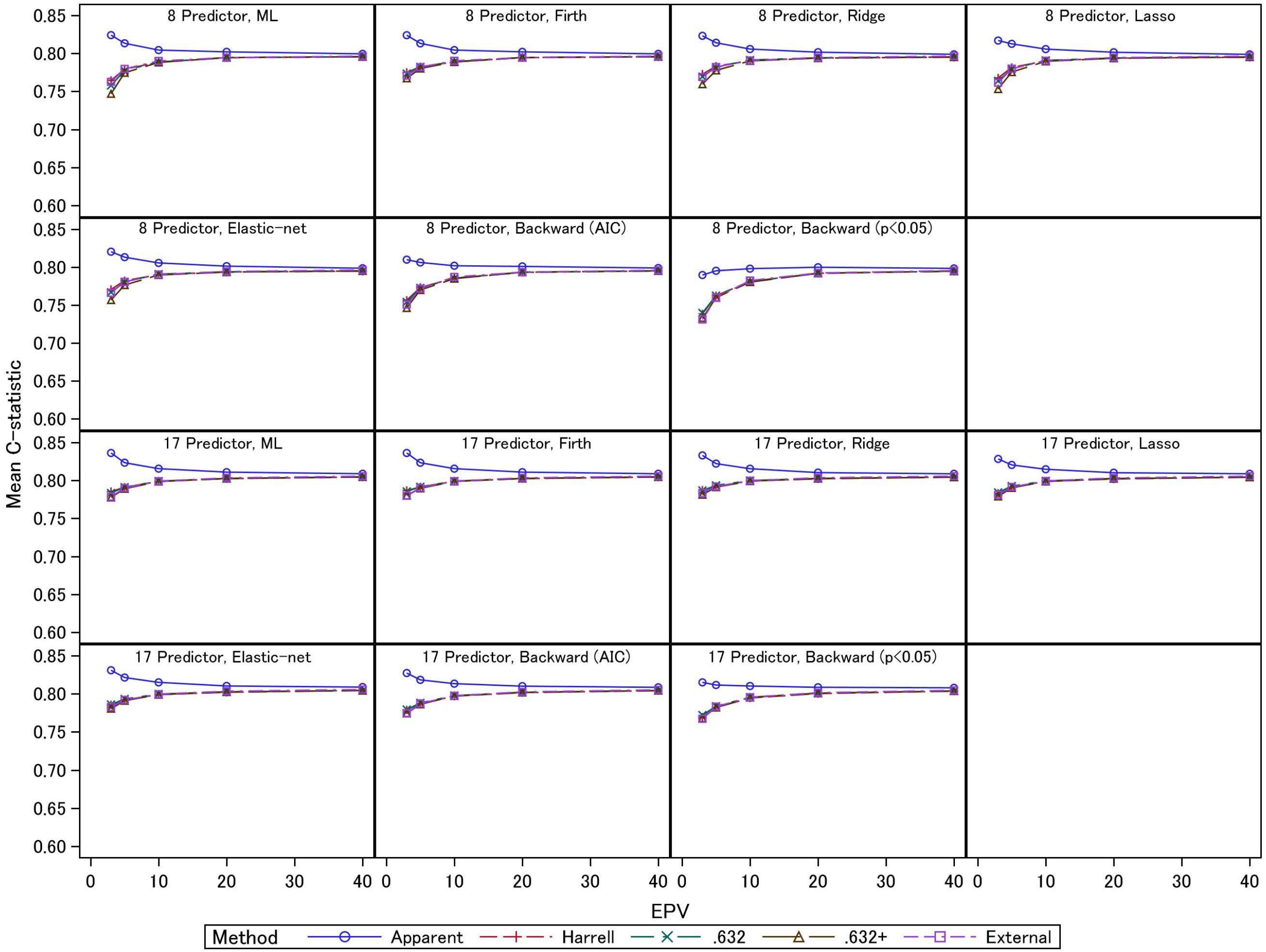

図 3.3.2.1-6 未調整、外部および各内的検証法の C 統計量（係数タイプ 1、イベント発生割合 0.0625）

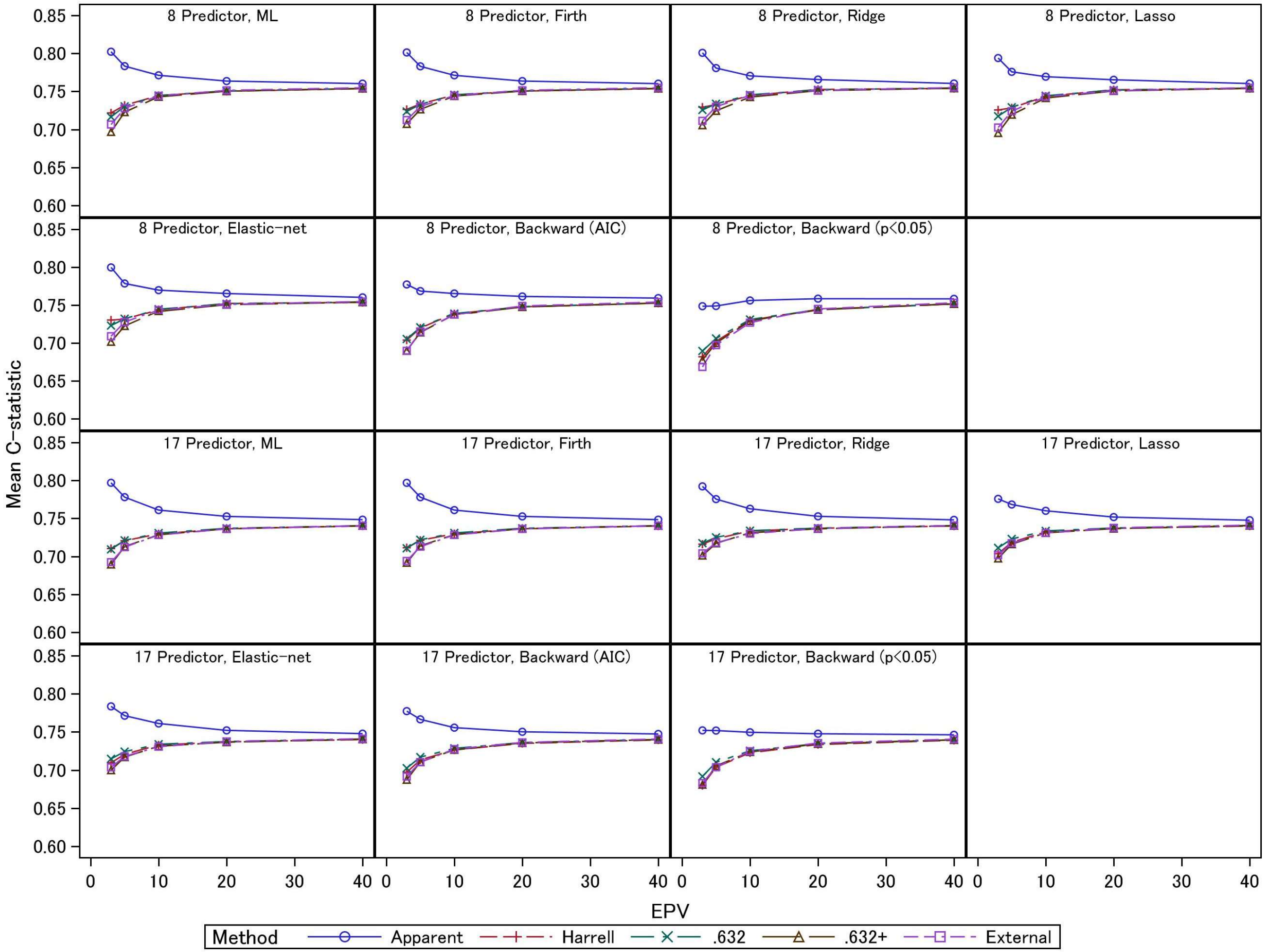

図 3.3.2.1-7 未調整、外部および各内的検証法の C 統計量（係数タイプ 2、イベント発生割合 0.25）

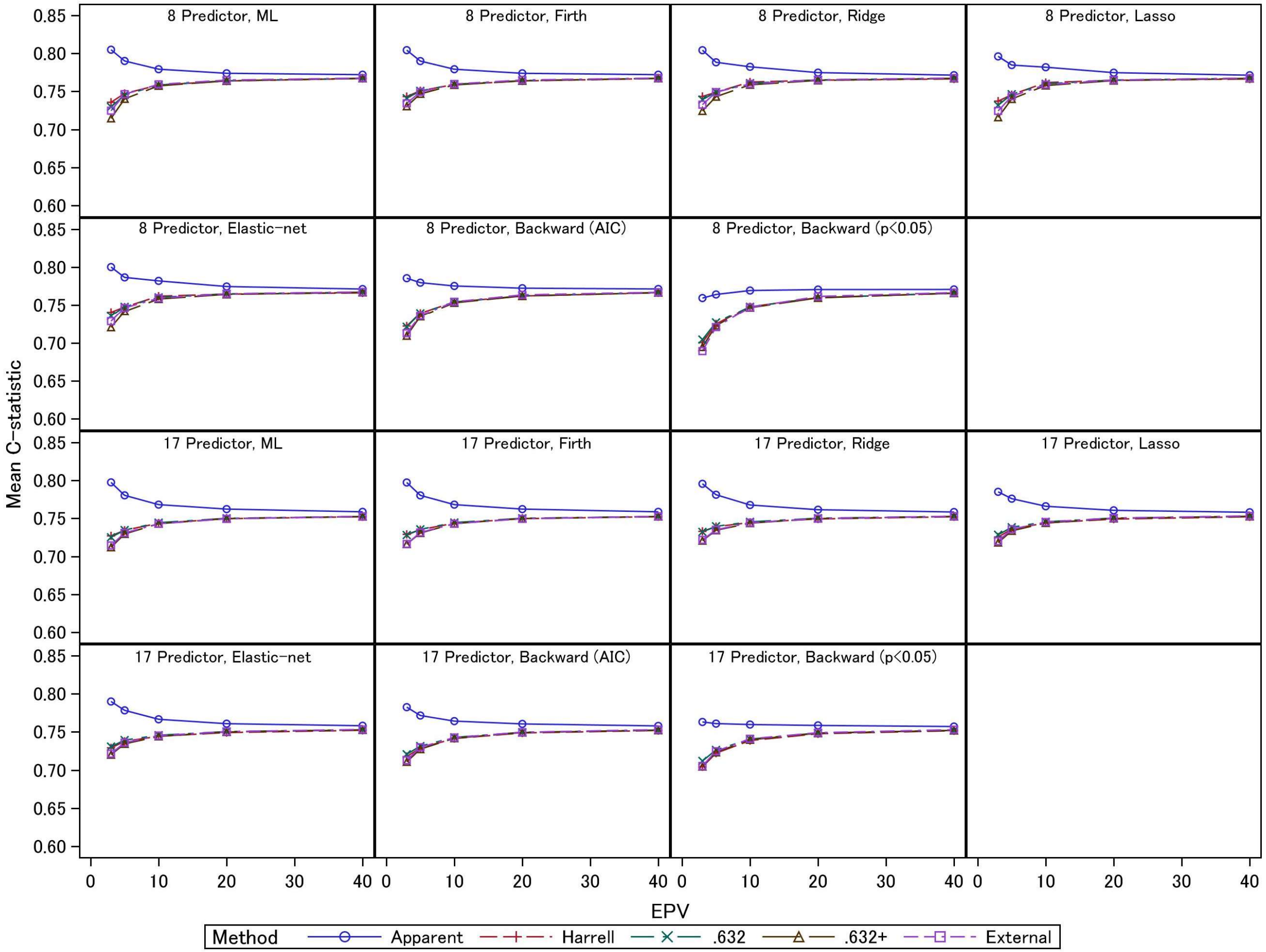

図 3.3.2.1-8 未調整、外部および各内的検証法の C 統計量（係数タイプ 2、イベント発生割合 0.125）

3.3.2.2 未調整および各内的検証法の C 統計量のバイアスの結果

係数タイプ 2 かつイベント発生割合 0.5 および 0.0625 における未調整および各内的検証法の C 統計量のバイアスを図 3.3.2.2-1 および図 3.3.2.2-2 に示した。なお、すべてのシナリオを通しての各 C 統計量のバイアスのモンテカルロ標準誤差の最大値は 0.0024 であった。

EPV が 3~5 では、未調整の C 統計量は、ブートストラップ法に基づく内的検証法の C 統計量よりも大きな過大評価のバイアス（イベント発生割合 0.5 で 0.07-0.16、イベント発生割合 0.0625 で 0.03-0.07）を示した。特に、EPV が小さいシナリオほど、過大評価のバイアスは大きくなった。同じ EPV のシナリオでは、イベント発生割合が小さいほど（すなわち、サンプルサイズが大きいほど）、過大評価のバイアスは小さくなった。EPV = 3 かつイベント発生割合 0.5 のもとでの 17 変数モデルでは、最尤法および Firth 法の未調整の C 統計量の過大評価のバイアス（0.16）は、他のモデル構築法と比較して大きかった。Ridge 回帰および AIC によるステップワイズ法の過大評価のバイアス（0.13）は、Elastic-net 回帰および Lasso 回帰の過大評価のバイアス（0.11-0.12）よりも大きかった。$P < 0.05$ の基準によるステップワイズ法は、最小の過大評価のバイアス（0.10）を示した。8 変数モデルでは、最尤法は他のモデル構築法と比較して、大きな過大評価のバイアス（0.14）を示した。縮小推定法の過大評価のバイアスは同程度（0.13）であった。ステップワイズ法は、他のモデル構築法と比較して、小さな過大評価のバイアス（0.11-0.12）を示したが、上述したように外部の C 統計量も小さかった。モデル構築法間の過大評価のバイアスの差は、EPV が大きくなるにつれて小さくなった。未調整の C 統計量のバイアスについて、イベント発生割合 0.0625 のシナリオにおいても、同様の傾向が認められた。

EPV ≥ 20 の全てのシナリオで、3 つのブートストラップ法に基づく内的検証法の C 統計量は同等であり、いずれにもバイアスが認められなかった。また、EPV = 10 の全てのシナリオにおいて、各内的検証法の C 統計量のバイアスの絶対値は、いずれも 0.01 未満であった。従来の最尤法の結果は、Steyerberg et al. [4]によって報告された結果と一致しており、縮小推定法およびステップワイズ法についても、同様の結果が得られることが確認できた。また、EPV ≥ 5 において、イベント発生割合 0.0625 の場合、サンプルサイズが相対的に大きいため、いずれの内的検証法の C 統計量にもバイアスは認められなかった。EPV = 3 のもとでの 8 変数モデルでは、.632+法は若干の過小評価のバイアスを示し、最尤法での過小評価のバイアス（0.02）が最も大きかった。Ridge 回帰、Lasso 回帰および Elastic-net 回帰での.632+法の過小評価のバイアスは 0.01 であった。Firth 法およびステップワイズ法では、.632+法の過小評価のバイアスは 0.01 未満であった。Harrell 法および.632 法は同様の傾向であり、0.01 以下の過大評価のバイアスを示した。17 変数モデルでは、.632+法の過小評価のバイアスは 0.01 未満であったが、.632+法は一般的に過小評価のバイアスを示す傾向があった。Harrell 法および.632 法の過大評価のバイアスは 0.01 未満であった。イベント発生割合 0.5 のもとでの 8 変数モデルでは、Harrell 法および.632 法の過大評価のバイアスは顕著に大きく、EPV が 3～5 において、0.03-0.04 のバイアスが認められた。.632+法は、最尤法および Firth 法で過小評価のバイアス（EPV = 3 で 0.01）を示したが、Ridge 回帰、Lasso 回帰および Elastic-net 回帰では、ほとんどバイアスを示さなかった。ステップワイズ法に関しては、AIC の場合は、ほとんどバイアスが認められなかったのに対して、$P < 0.05$ の基準では過大評価のバイアス（0.02）が認められた。17 変数モデルにおいても同様の傾向が認められた。Harrell 法および.632 法は、EPV = 3 において、0.02-0.04 の過大評価のバイアスを示した。この過大評価のバイアスは、最尤法、Firth 法および

Ridge 回帰では同程度であった。しかしながら、Lasso 回帰、Elastic-net 回帰およびAIC を用いたステップワイズ法では、Harrell 法と比べて、.632 法の過大評価のバイアスが大きかった。$P < 0.05$ の基準を用いたステップワイズ法では、.632 法は過大評価のバイアスを示したが、Harrell 法はバイアスを示さなかった。.632+法は、最尤法、Firth 法およびAIC を用いたステップワイズ法で、0.01 未満の過小評価のバイアスを示したが、Ridge 回帰、Lasso 回帰およびElastic-net 回帰および $P < 0.05$ の基準を用いたステップワイズ法では、ほとんどバイアスを示さなかった。

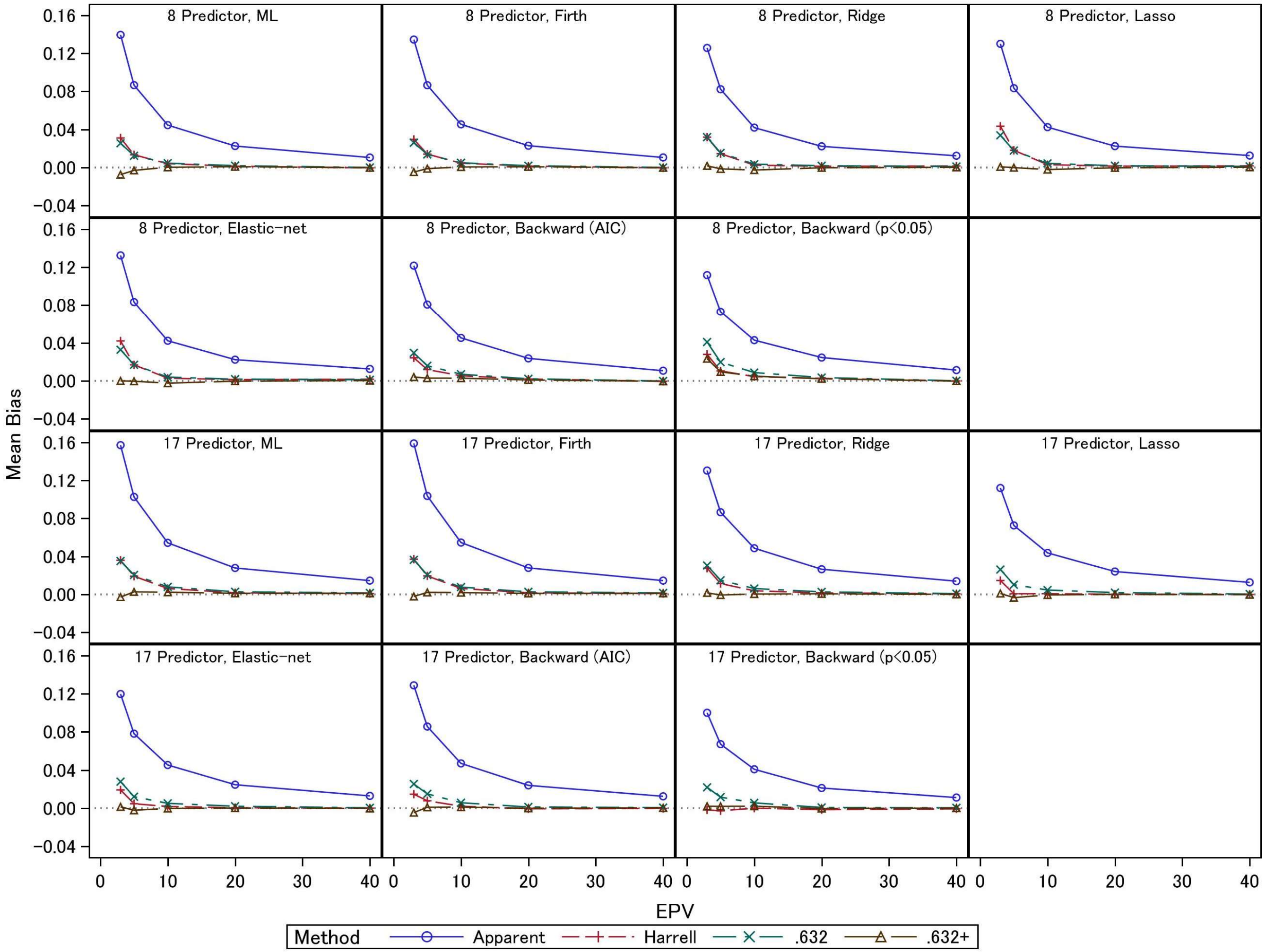

図 3.3.2.2-1 未調整および各内的検証法の C 統計量のバイアス（係数タイプ 2、イベント発生割合 0.5）

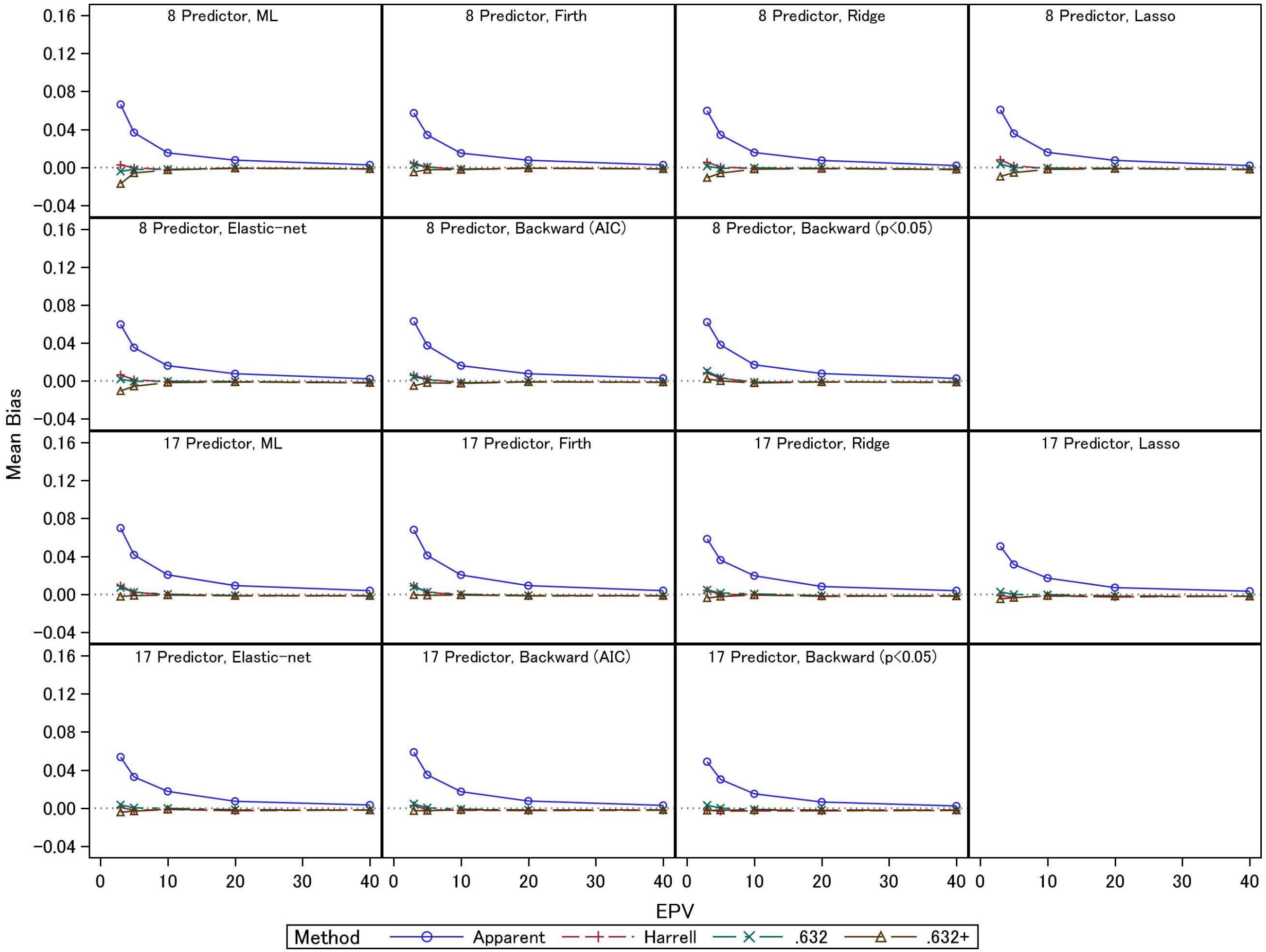

図 3.3.2.2-2 未調整および各内的検証法の C 統計量のバイアス（係数タイプ 2、イベント発生割合 0.0625）

係数タイプ 1 のイベント発生割合 0.5、0.25、0.125、0.0625 および係数タイプ 2 のイベント発生割合 0.25、0.125 における未調整および各内的検証法の C 統計量のバイアスを図 3.3.2.2-3〜図 3.3.2.2-8 に示した。

係数タイプ 2 のイベント発生割合 0.25 および 0.125 において、各モデル構築法の未調整の C 統計量のバイアスは、イベント発生割合 0.5 および 0.0625 と同様の傾向であった。各内的検証法の C 統計量の比較に関しては、イベント発生割合 0.5 および 0.0625 の中間の傾向を示した。いずれのブートストラップ法に基づく推定量も、EPV が 5 以上では、顕著なバイアスは示さなかった。EPV = 3 では、イベント発生割合 0.125 のもとでの 8 変数モデルにおいて、.632+法の過小評価のバイアスは、最尤法では 0.01、他のモデル構築法では 0.01 未満であった。Harrell 法および.632 法は、約 0.01 以下の過大評価のバイアスを示した。17 変数モデルでは、.632+法には、ほとんどバイアスが認められなかった。Harrell 法および.632 法の過大評価のバイアスは、約 0.01 以下であった。イベント発生割合 0.25 のもとでの 8 変数モデルにおいて、.632+法は、一般的に過小評価のバイアス（0.01 以下）を示したが、$P < 0.05$ の基準を用いたステップワイズ法でのみ、0.01 の過大評価のバイアスを示した。Harrell 法および.632 法の過大評価のバイアスは 0.01-0.02 であった。17 変数モデルでは、.632+法はバイアスを示さなかった。Harrell 法および.632 法では、0.01-0.02 の過大評価のバイアスが認められた。過大評価のバイアスは、最尤法、Firth 法、Ridge 回帰では同程度であったが、Lasso 回帰、Elastic-net 回帰およびステップワイズ法では、.632 法の方が Harrell 法よりも大きかった。

ノイズ変数の少ない係数タイプ 1 のシナリオでは、未調整の C 統計量の過大評価のバイアスは、係数タイプ 2 のシナリオよりも小さくなった。3 つのブートストラップ法による推定量の比較に関しては、係数タイプ 1 のシナリオと係

数タイプ 2 のシナリオで同様の傾向であったが、バイアスの絶対値は係数タイプ 1 のシナリオの方が小さかった。

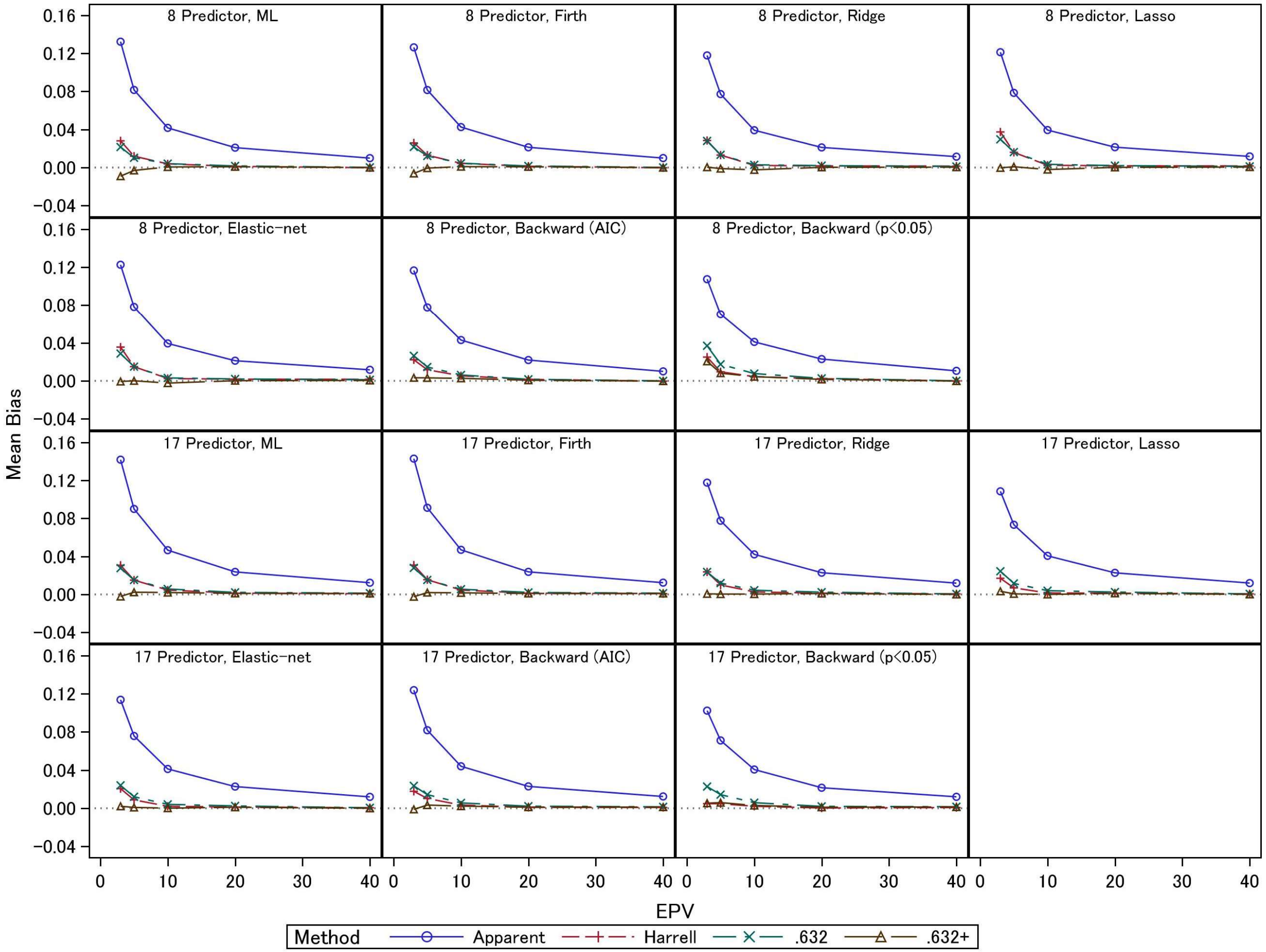

図 3.3.2.2-3 未調整および各内的検証法の C 統計量のバイアス（係数タイプ 1、イベント発生割合 0.5）

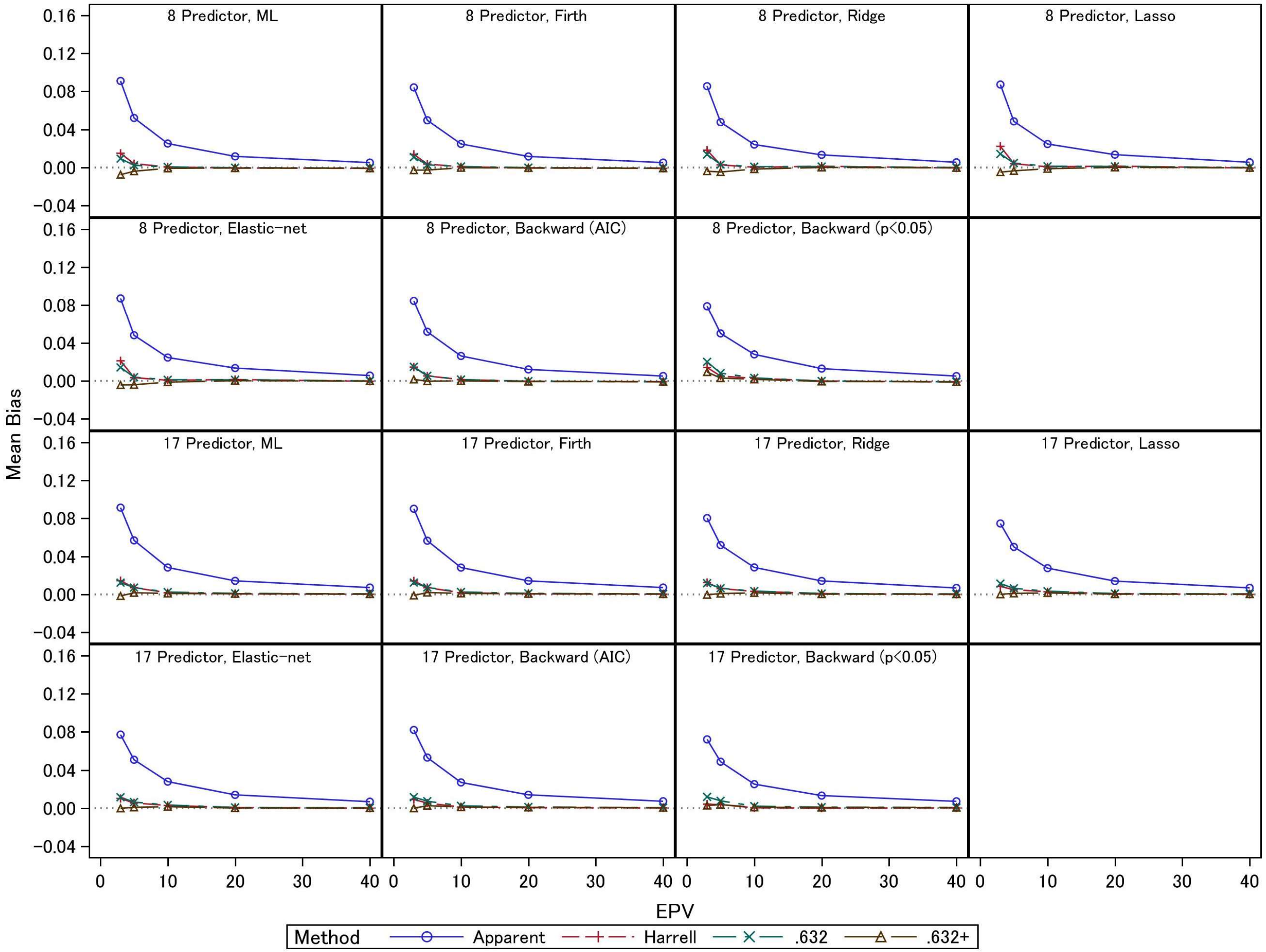

図 3.3.2.2-4 未調整および各内的検証法の C 統計量のバイアス（係数タイプ 1、イベント発生割合 0.25）

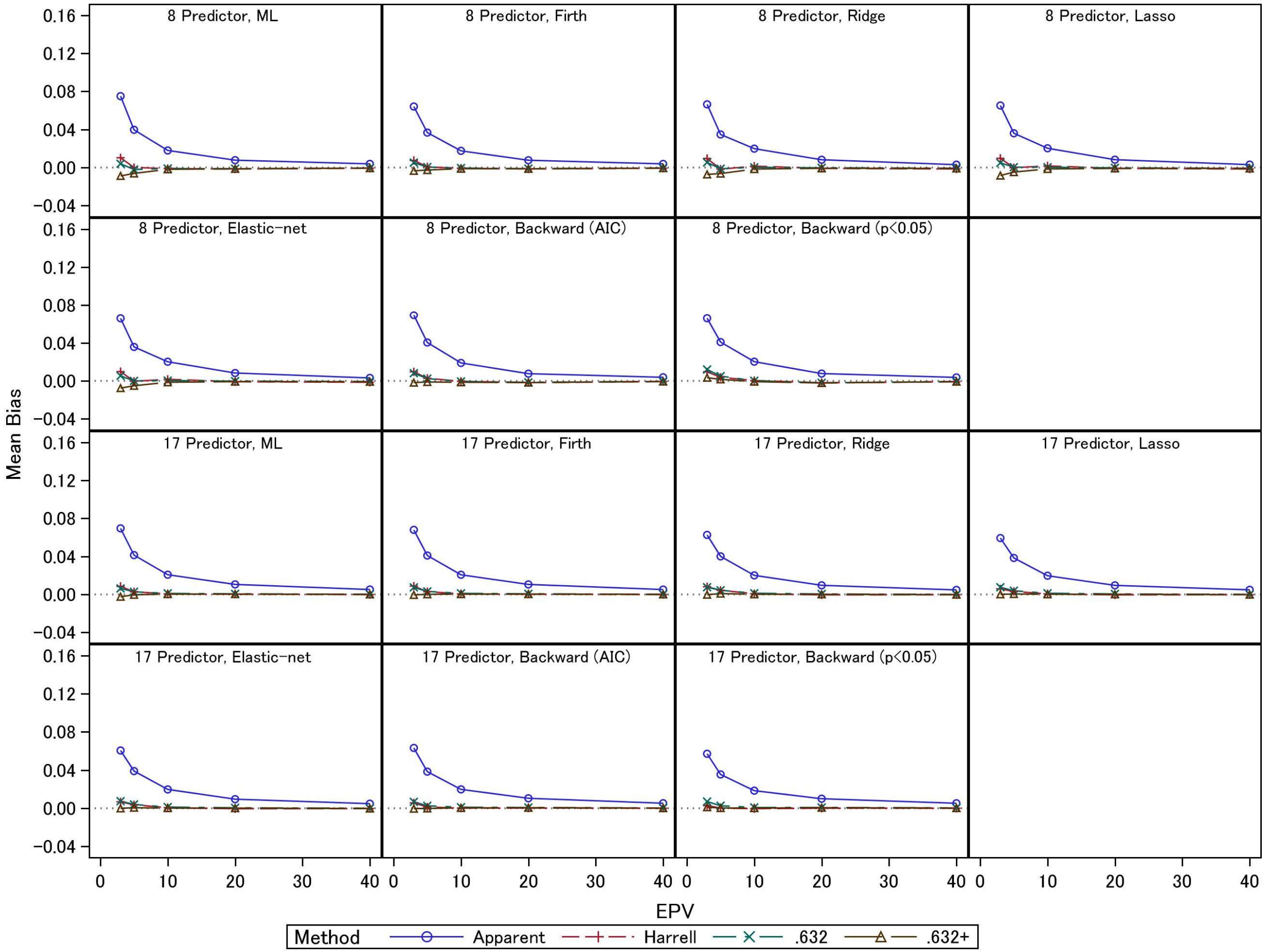

図 3.3.2.2-5 未調整および各内的検証法の C 統計量のバイアス（係数タイプ 1、イベント発生割合 0.125）

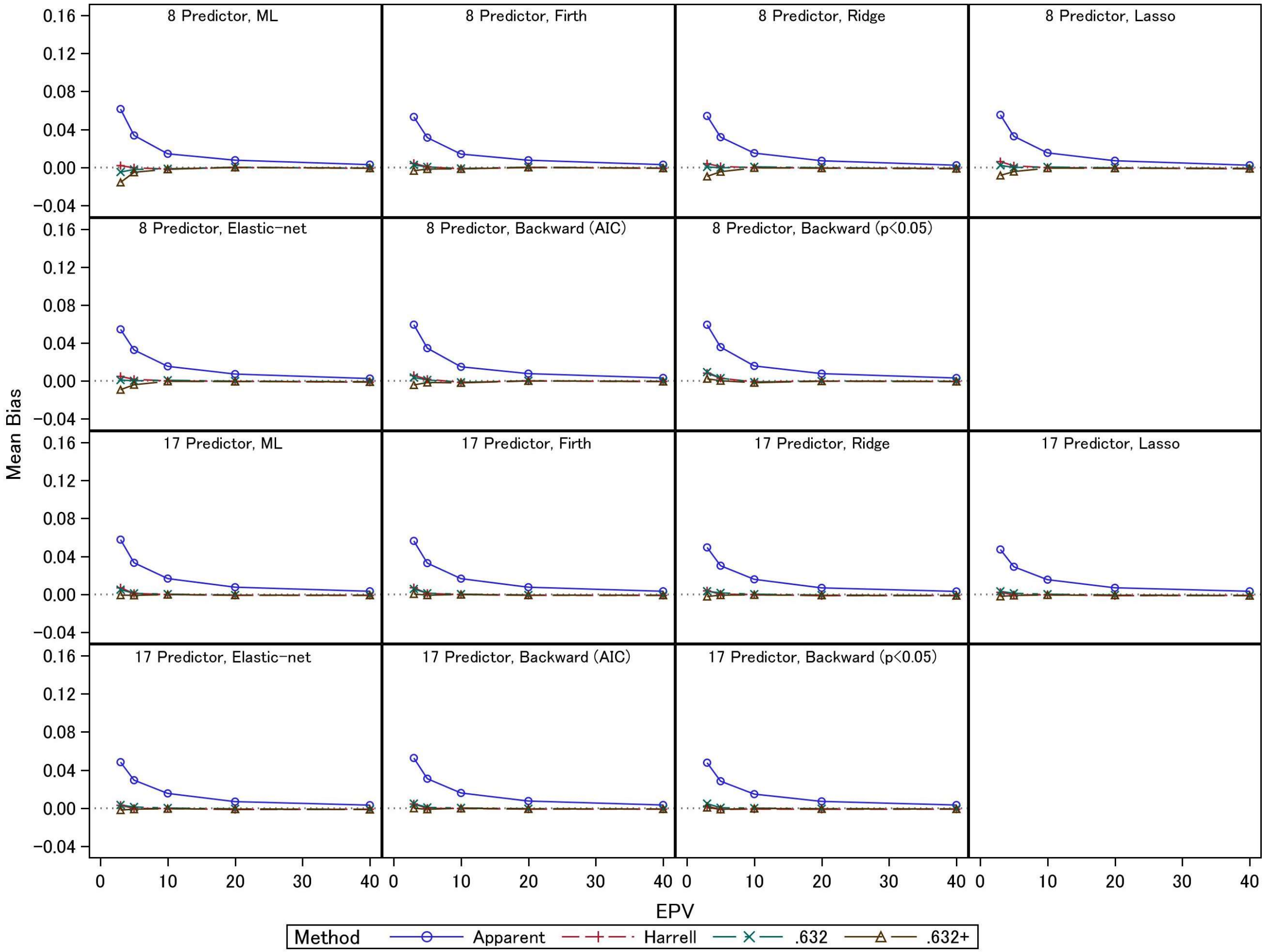

図 3.3.2.2-6 未調整および各内的検証法の C 統計量のバイアス（係数タイプ 1、イベント発生割合 0.0625）

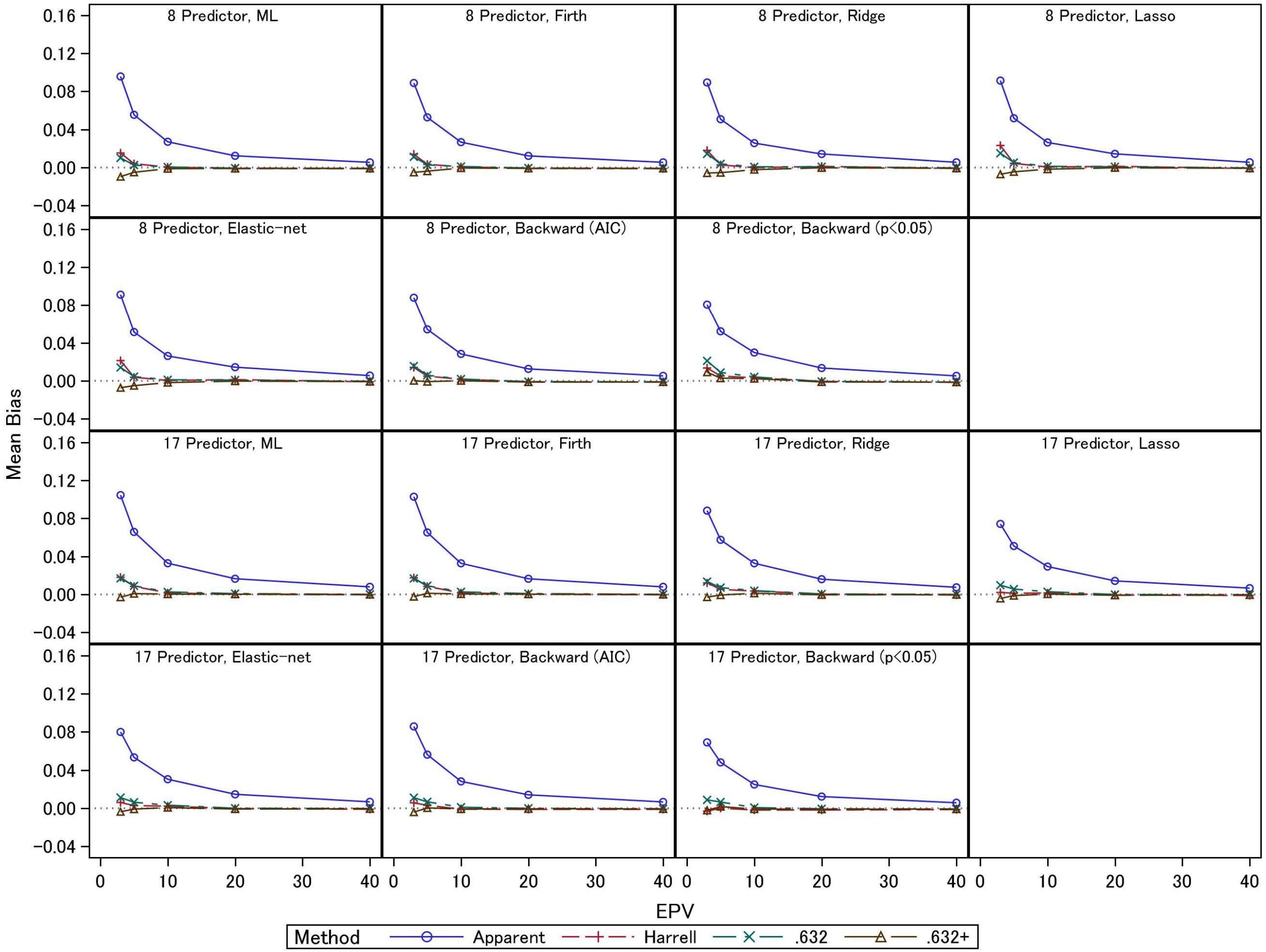

図 3.3.2.2-7 未調整および各内的検証法の C 統計量のバイアス（係数タイプ 2、イベント発生割合 0.25）

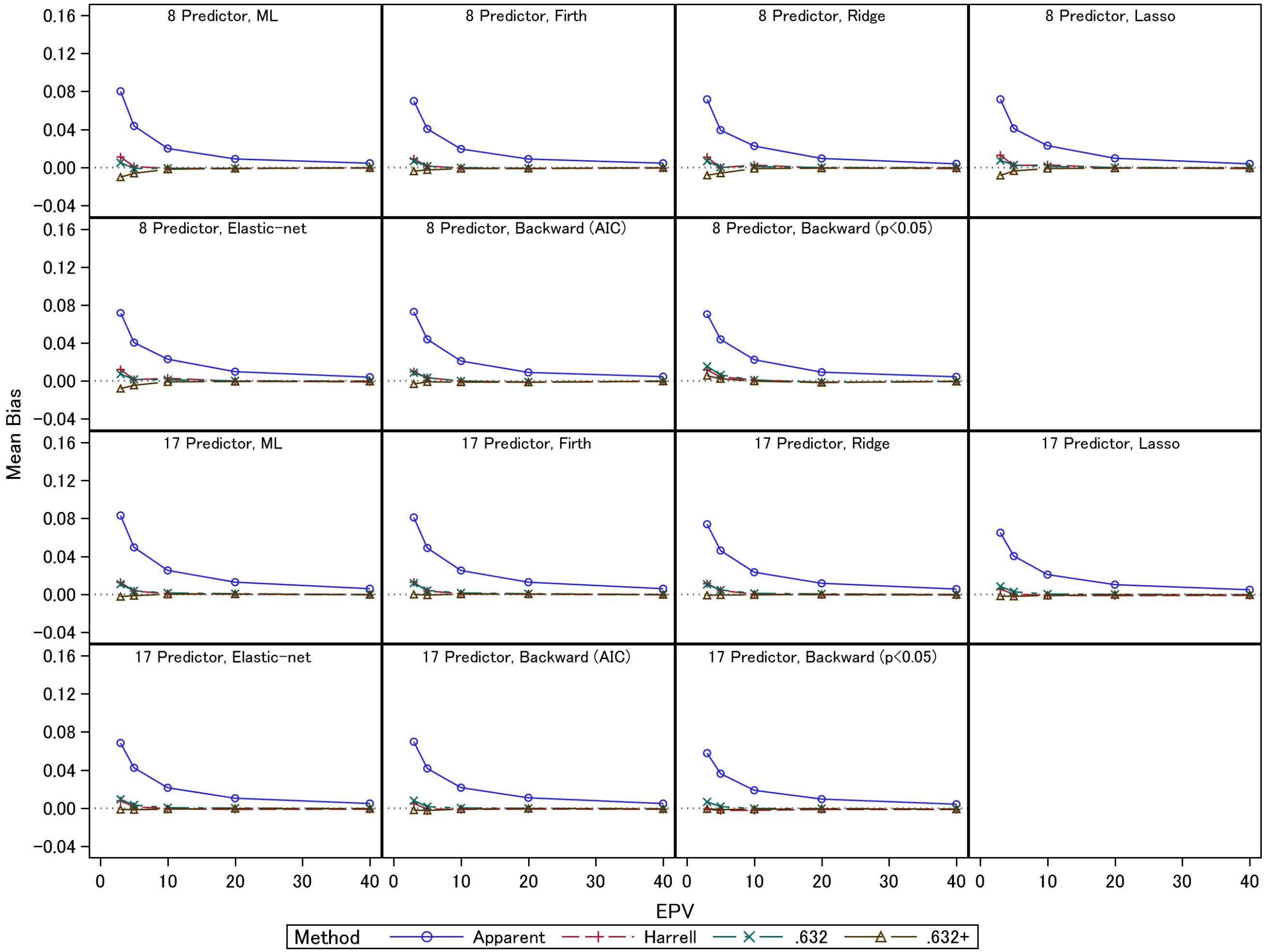

図 3.3.2.2-8 未調整および各内的検証法の C 統計量のバイアス（係数タイプ 2、イベント発生割合 0.125）

3.3.2.3 未調整および各内的検証法の C 統計量の RMSE の結果

係数タイプ 2 かつイベント発生割合 0.5 および 0.0625 における未調整および各内的検証法の C 統計量の RMSE を図 3.3.2.3-1 および図 3.3.2.3-2 に示した。なお、すべてのシナリオを通しての各 C 統計量の RMSE のモンテカルロ標準誤差の最大値は 0.0016 であった。

EPV＝3 および 5 において、未調整の C 統計量は、各内的検証法の C 統計量と比べて大きな RMSE（イベント発生割合 0.5 で 0.08-0.16、イベント発生割合 0.0625 で 0.04-0.08）を示した。この結果は、上述した小標本での未調整の C 統計量の大きなバイアスによって引き起こされた。3 つのブートストラップ法に基づく推定量の RMSE は、一般的に同程度であった。例外として、EPV が 3 および 5 かつイベント発生割合 0.5 の場合、8 変数モデルの Ridge 回帰、Lasso 回帰および Elastic-net 回帰では、.632+法の RMSE（0.07-0.10）は、他の 2 つの方法の RMSE（0.05-0.08）よりも大きかった。上述したように、これらのシナリオでは、.632+法のバイアスの絶対値は小さかったことから、この結果は、これらの推定量の標準誤差を反映していた。なお、17 変数モデルでは、3 つのブートストラップ法に基づく推定量の RMSE は同程度であった。EPV が 3 かつイベント発生割合 0.0625 の場合でも、8 変数モデルの Ridge 回帰、Lasso 回帰および Elastic-net 回帰では、.632+法の RMSE（0.06-0.07）は、他の 2 つの方法の RMSE（0.05）と比較して大きかった。また、.632+法の過小評価のバイアスが大きかった最尤法においても、.632+法の RMSE（0.07）は、他の 2 つの方法の RMSE（0.06）よりも若干大きかった。上記以外の EPV が小さいシナリオにおいて、.632+法のバイアスが他の 2 つの方法よりも小さいシナリオが多くあったが、そのようなシナリオにおいても、.632+法の RMSE は、他の 2 つの方法と同程度であった。この結果は、EPV が小さいシナリオにおいて、.632+法の標準誤差が、他の 2 つの方法の標準誤差よりも大きかったことを示唆した。

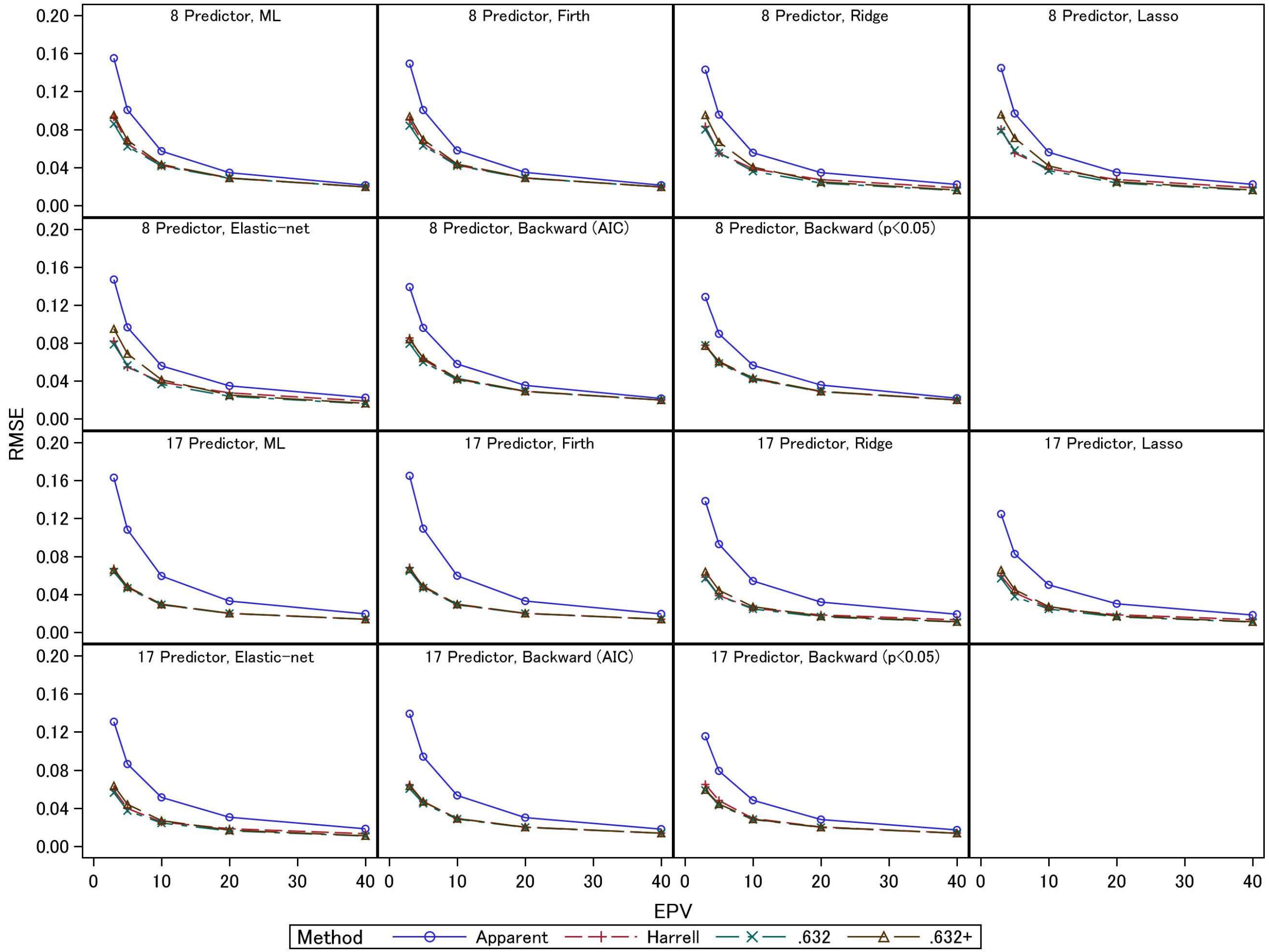

図 3.3.2.3-1 未調整および各内的検証法の C 統計量の RMSE（係数タイプ 2、イベント発生割合 0.5）

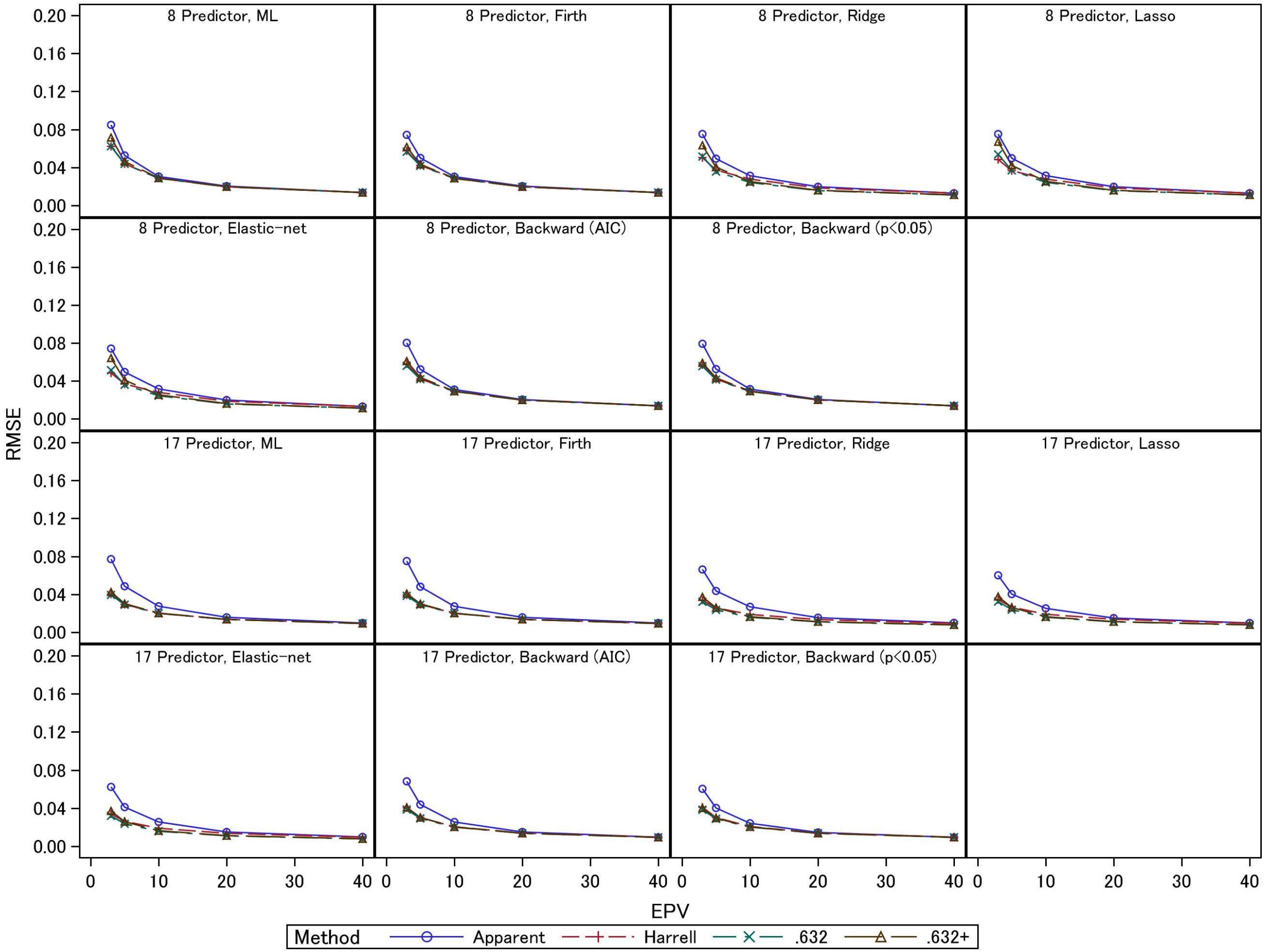

図 3.3.2.3-2 未調整および各内的検証法の C 統計量の RMSE（係数タイプ 2、イベント発生割合 0.0625）

係数タイプ 1 のイベント発生割合 0.5、0.25、0.125、0.0625 および係数タイプ 2 のイベント発生割合 0.25、0.125 における未調整および各内的検証法の C 統計量の RMSE を図 3.3.2.3-3～図 3.3.2.3-8 に示した。

係数タイプ 2 かつイベント発生割合 0.25 および 0.125 における 3 つのブートストラップ法に基づく推定量の RMSE の比較では、イベント発生割合 0.5 および 0.0625 と同様の傾向が認められた。8 変数モデルの Ridge 回帰、Lasso 回帰および Elastic-net 回帰において、.632+法の RMSE（EPV が 3～5 で 0.05-0.08；イベント発生割合 0.25、0.04-0.07；イベント発生割合 0.125）は、他の 2 つの方法の RMSE（EPV が 3～5 で 0.05-0.06；イベント発生割合 0.25、0.04-0.05；イベント発生割合 0.125）よりも大きかったが、他のモデル構築法では、各推定量の RMSE は同程度であった。17 変数モデルでは、各推定量の RMSE に大きな違いは認められなかった。

係数タイプ 1 のシナリオでの RMSE は、一般的に係数タイプ 2 のシナリオでの RMSE よりも小さくなった。3 つのブートストラップ法に基づく推定量の比較は、係数タイプ 2 のシナリオと同様の傾向であり、小標本で 8 変数モデルの Ridge 回帰、Lasso 回帰および Elastic-net 回帰を用いた場合を除いて、これらの推定量の RMSE は同程度であった。

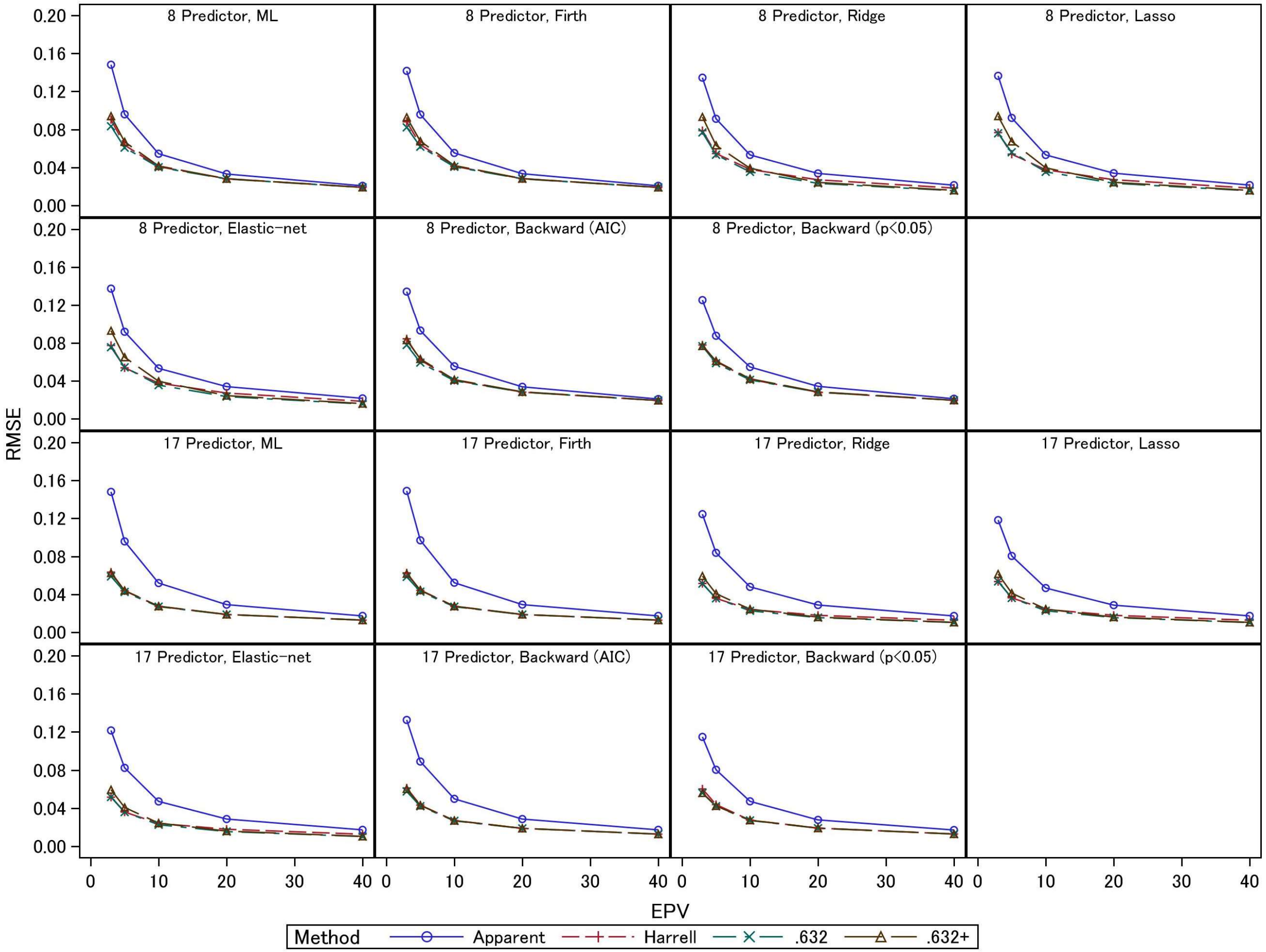

図 3.3.2.3-3 未調整および各内的検証法の C 統計量の RMSE（係数タイプ 1、イベント発生割合 0.5）

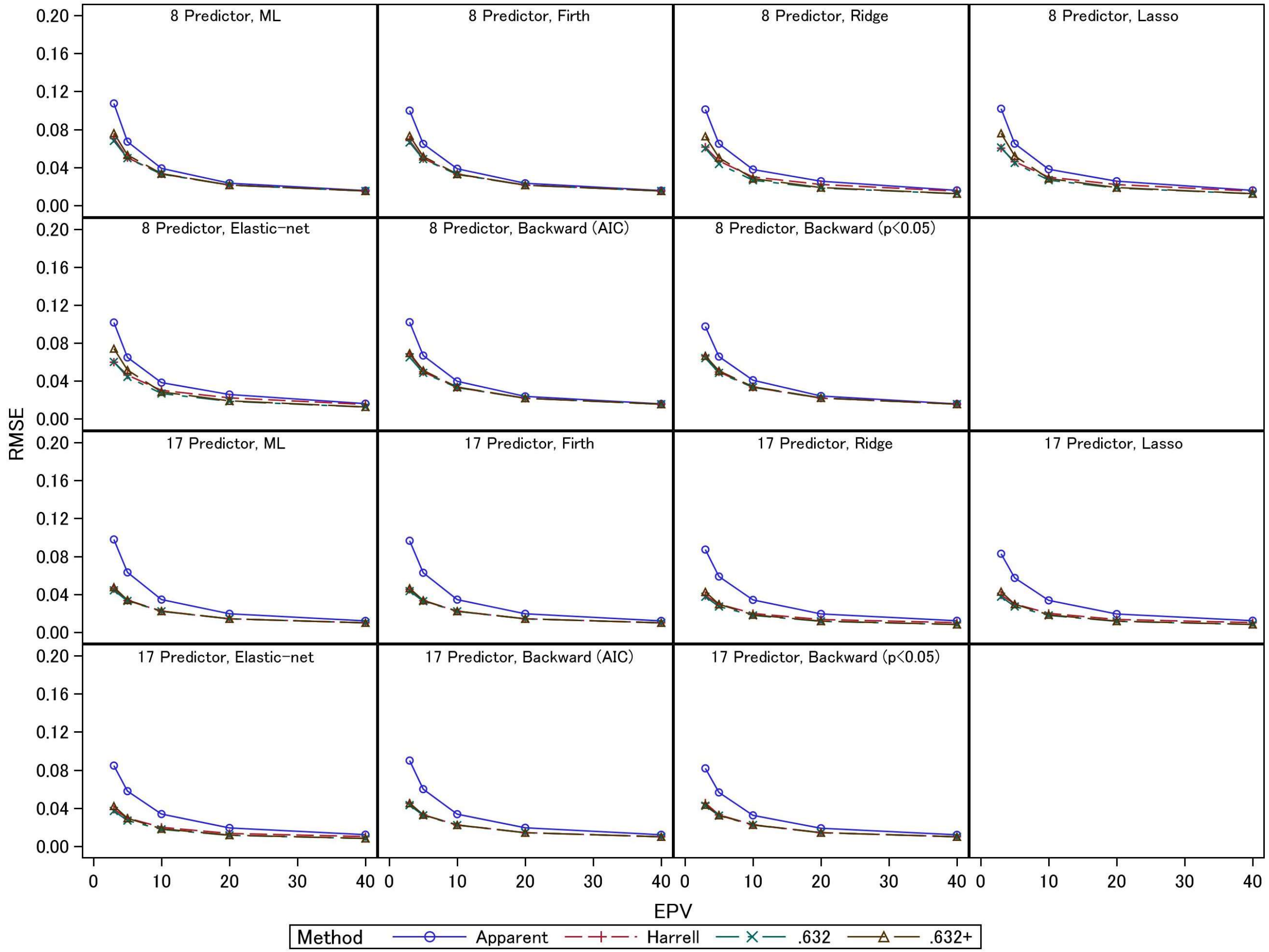

図 3.3.2.3-4 未調整および各内的検証法の C 統計量の RMSE（係数タイプ 1、イベント発生割合 0.25）

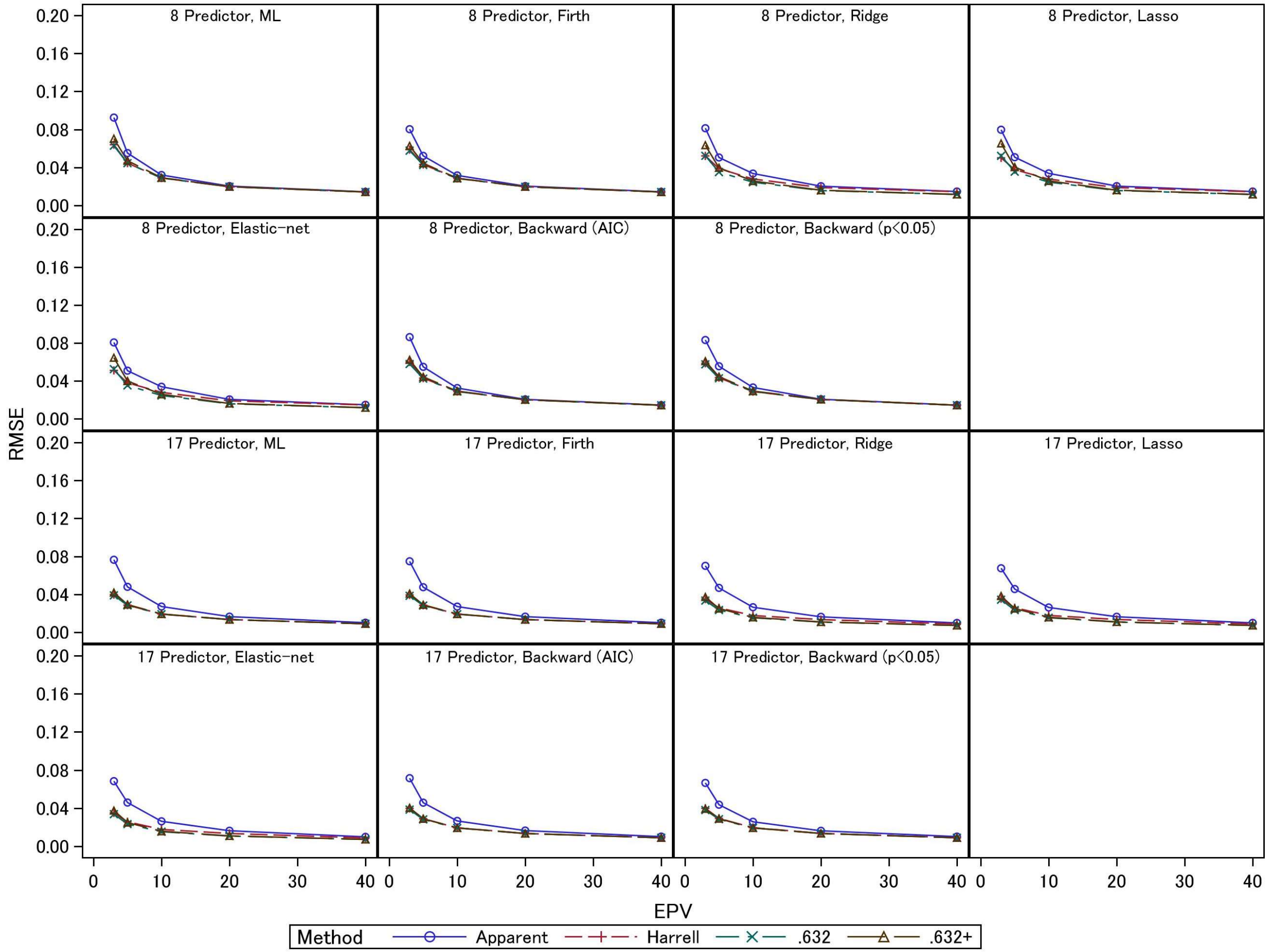

図 3.3.2.3-5 未調整および各内的検証法の C 統計量の RMSE（係数タイプ 1、イベント発生割合 0.125）

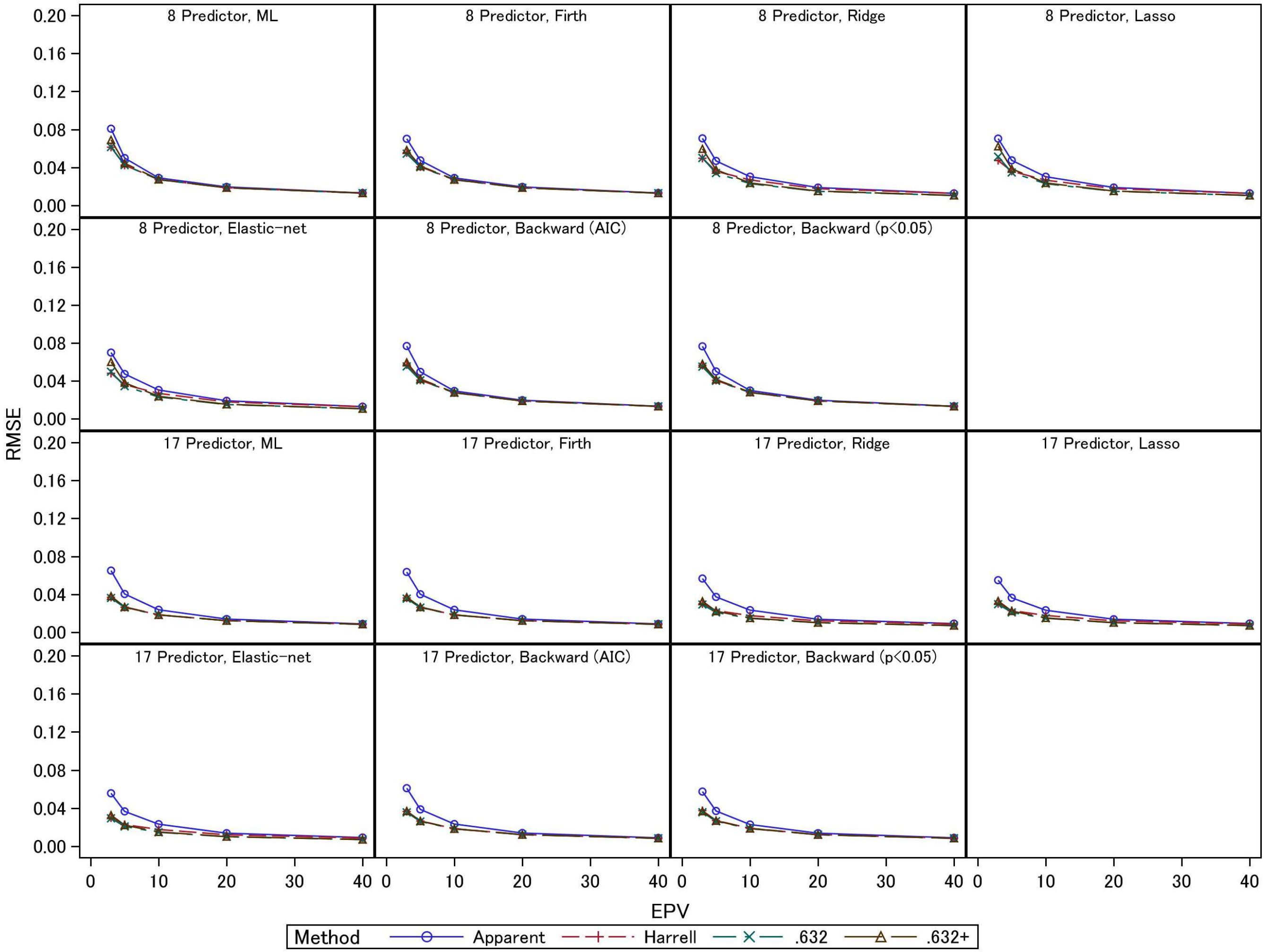

図 3.3.2.3-6 未調整および各内的検証法の C 統計量の RMSE（係数タイプ 1、イベント発生割合 0.0625）

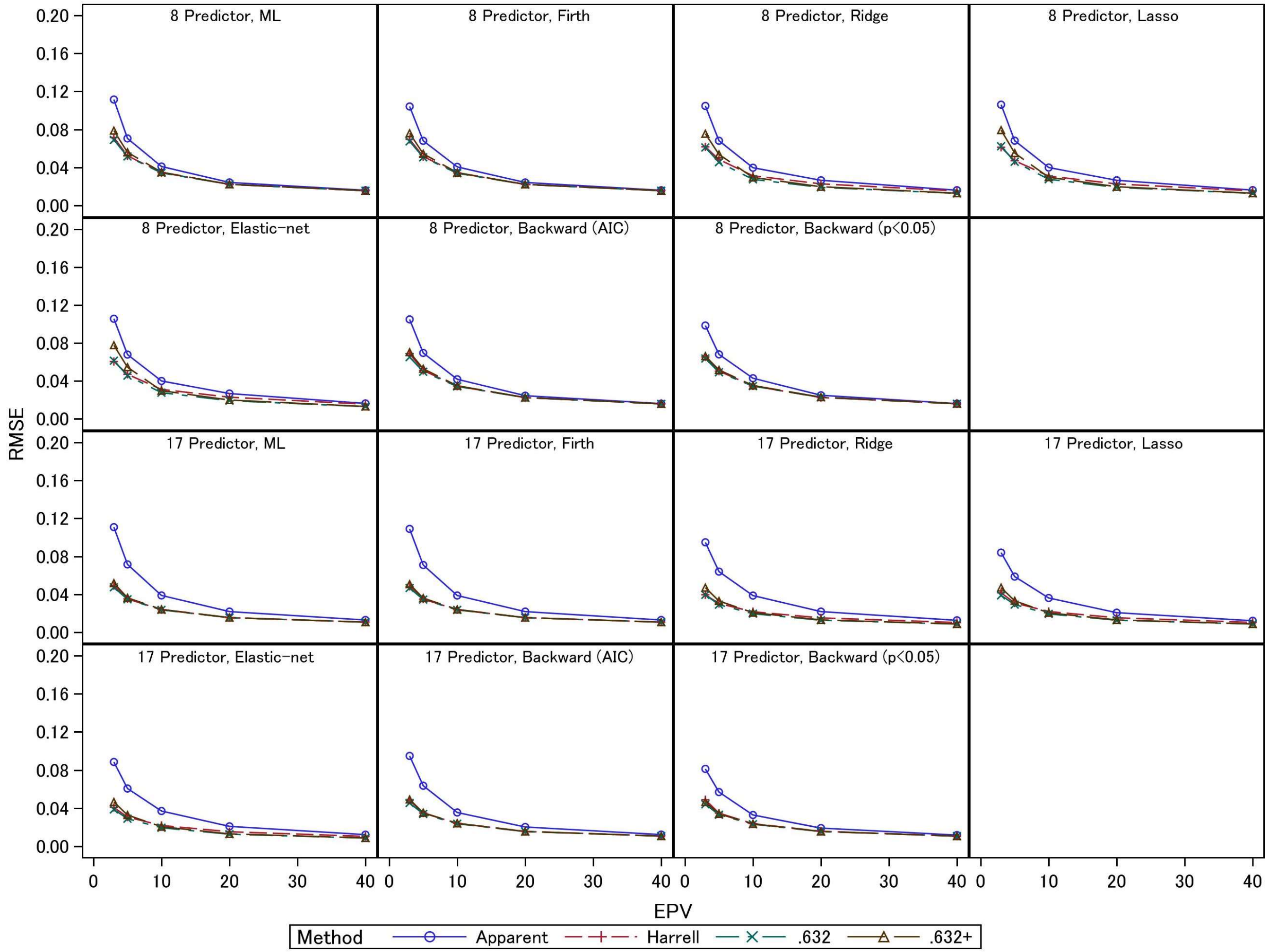

図 3.3.2.3-7 未調整および各内的検証法の C 統計量の RMSE（係数タイプ 2、イベント発生割合 0.25）

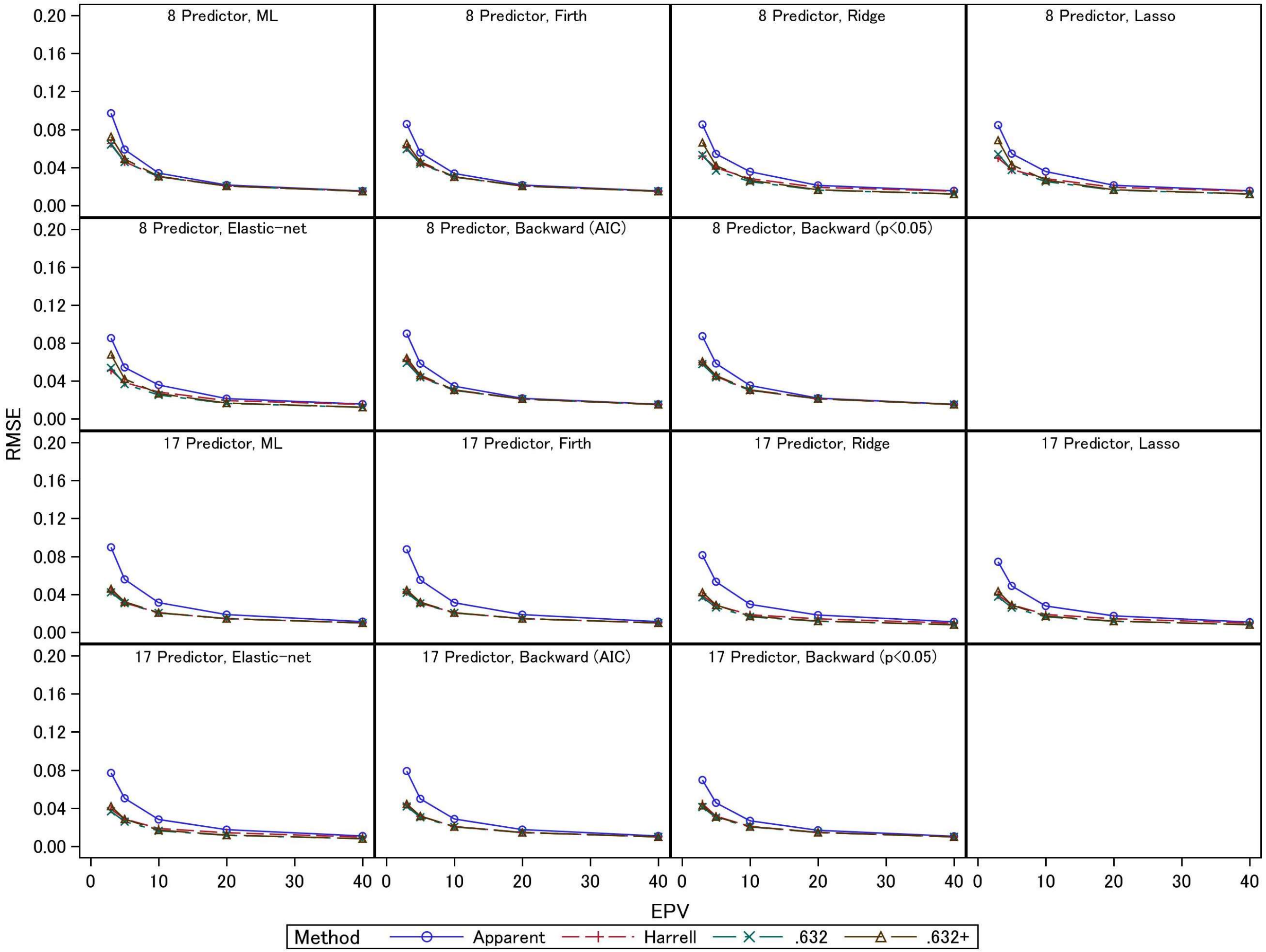

図 3.3.2.3-8 未調整および各内的検証法の C 統計量の RMSE（係数タイプ 2、イベント発生割合 0.125）

3.3.2.4 多変量歪正規分布を用いたシミュレーション実験

多変量歪正規分布を用いた連続変数の分布の歪みに対する感度分析の結果を以下に示した。感度分析は、最尤法で構築した予測モデルについてのみ行った。未調整、外部および各内的検証法の C 統計量の平均値を図 3.3.2.4-1 および図 3.3.2.4-2（係数タイプ 1 および係数タイプ 2、以下同順）に示した。未調整および各内的検証法の C 統計量のバイアスを図 3.3.2.4-3 および図 3.3.2.4-4 に示した。未調整および各内的検証法の C 統計量の RMSE を図 3.3.2.4-5 および図 3.3.2.4-6 に示した。

GUSTO-I 試験 Western データセットから推定した身長、体重および年齢の多変量歪正規分布の歪度パラメータは、−1.1、3.2 および 0.0 であり、身長および体重の分布が歪んでいることが示唆された。多変量正規分布および多変量歪正規分布に基づくシミュレーション実験の結果は、全てのシナリオで同等であったことから、連続変数の分布の歪みは、シミュレーション実験の結果に影響しなかった。

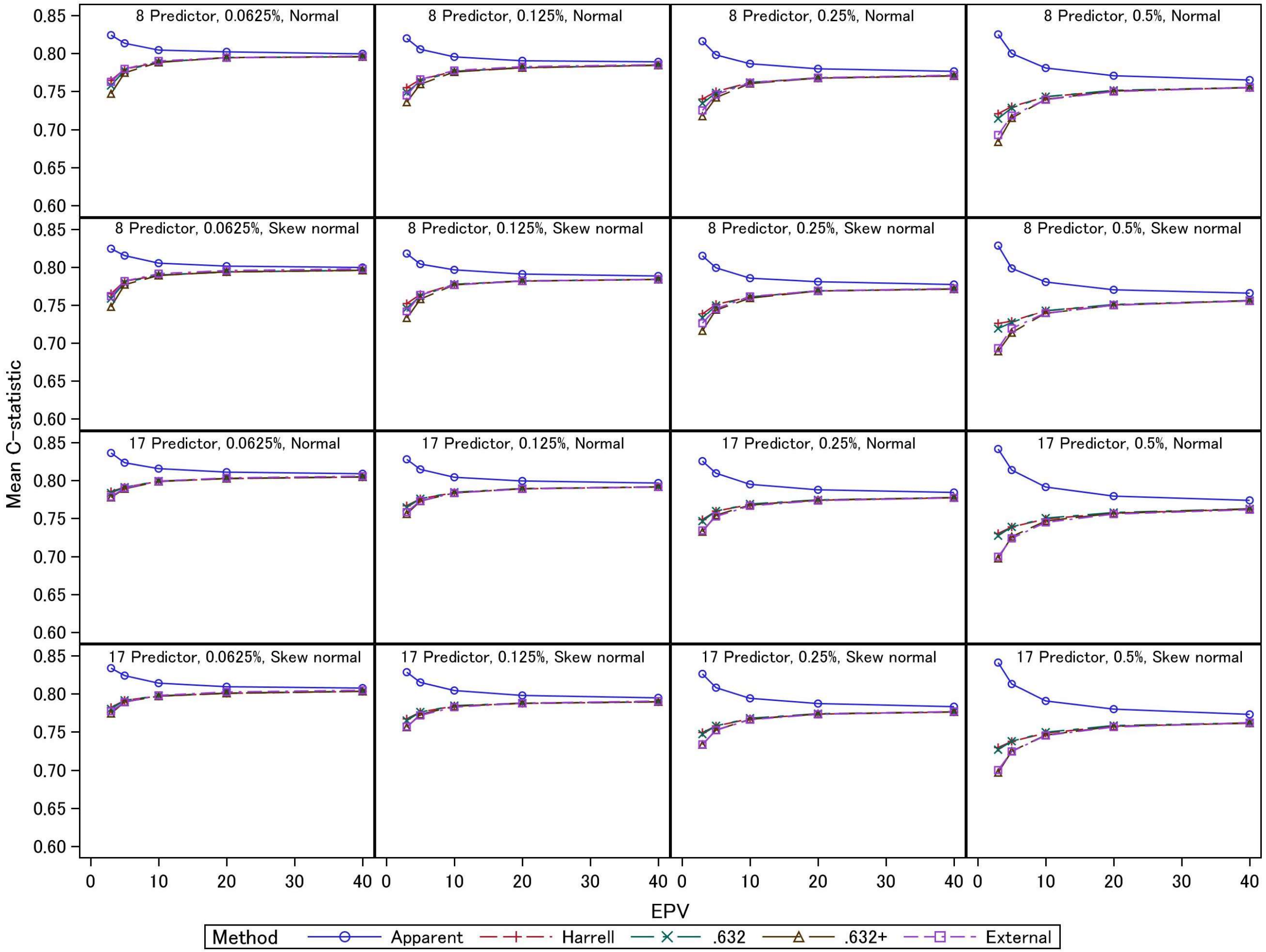

図 3.3.2.4-1 多変量正規および歪正規分布を用いた場合の未調整、外部および各内的検証法の C 統計量（係数タイプ 1）

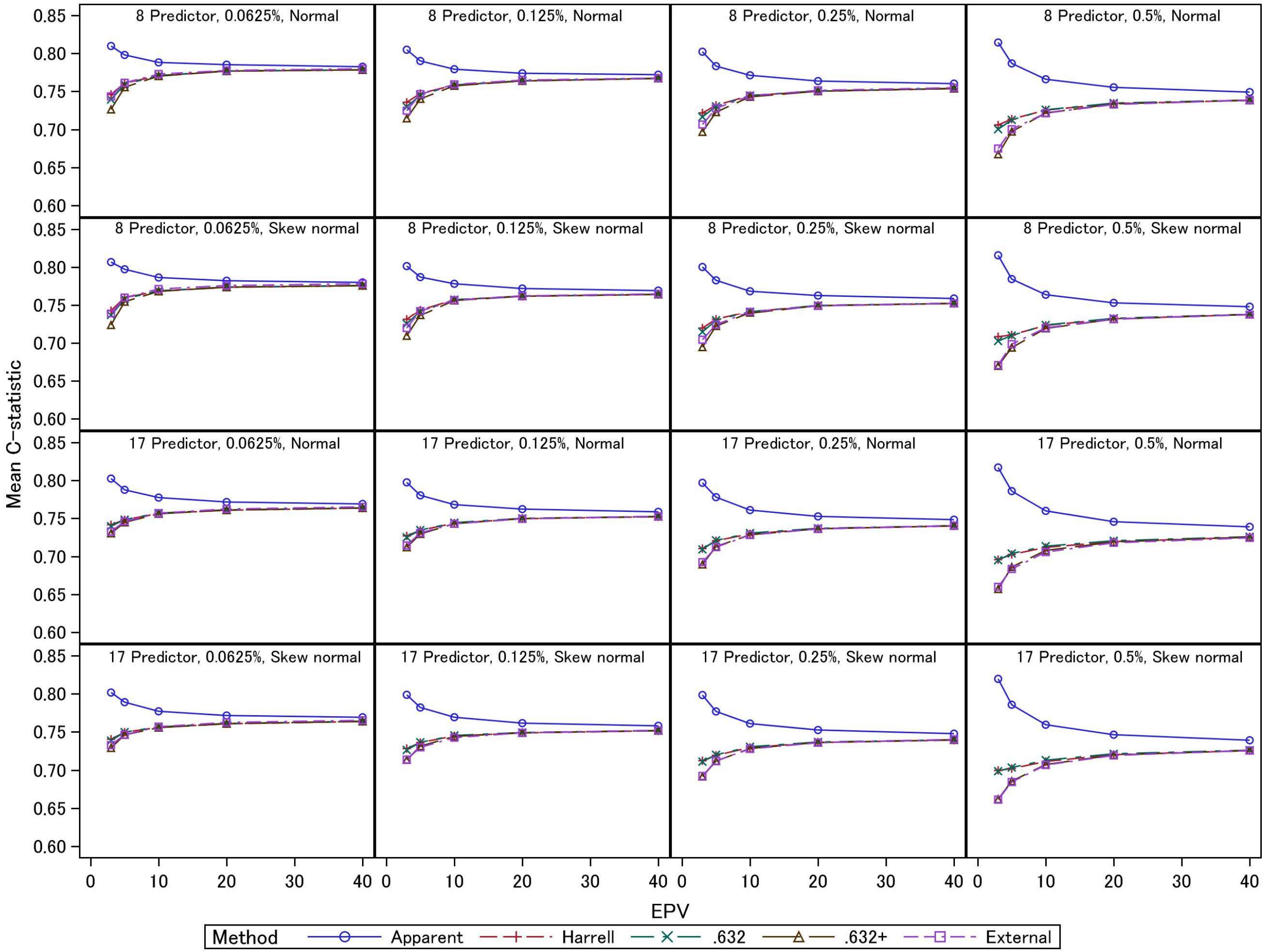

図 3.3.2.4-2 多変量正規および歪正規分布を用いた場合の未調整、外部および各内的検証法の C 統計量（係数タイプ 2）

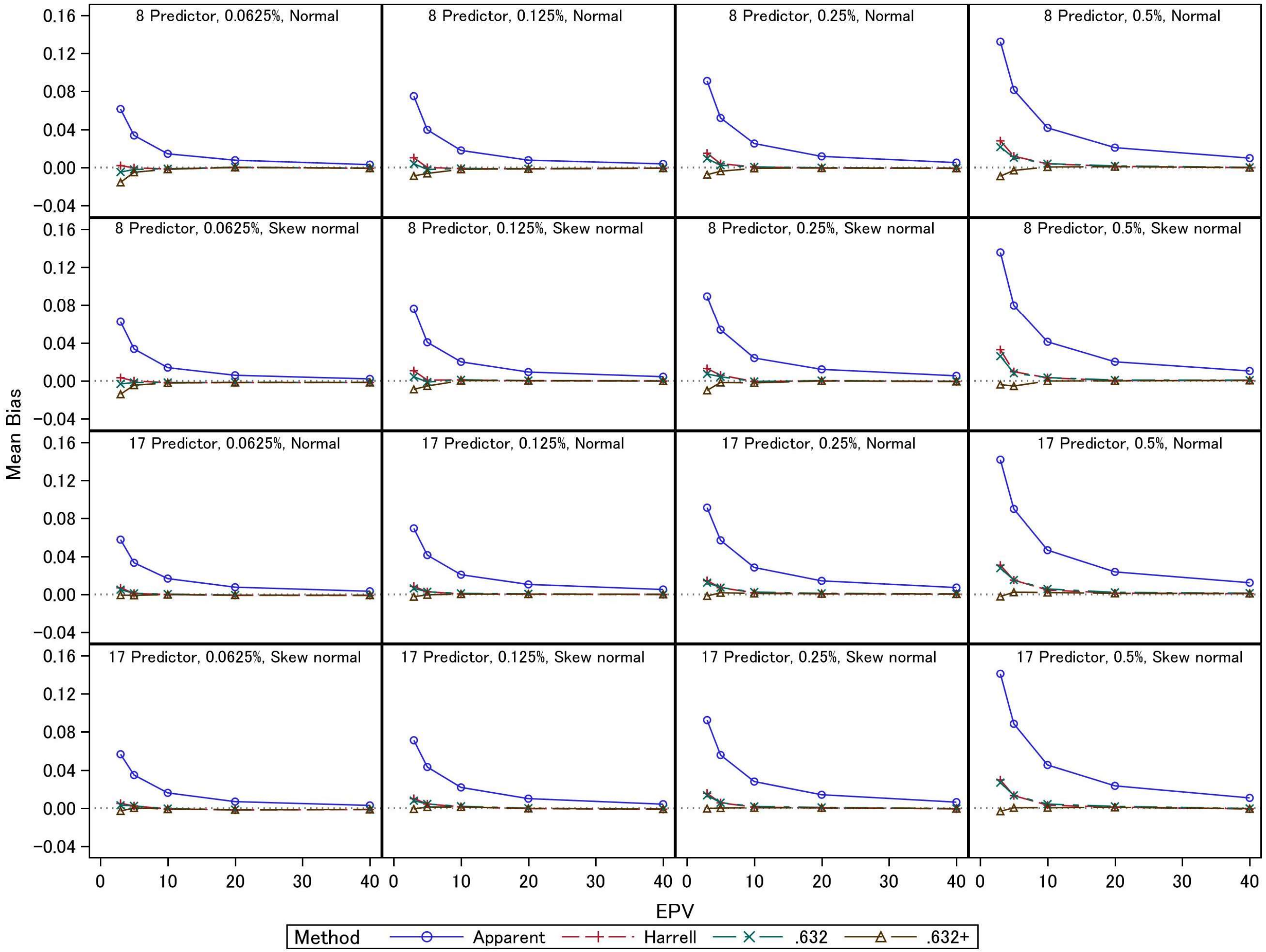

図 3.3.2.4-3 多変量正規および歪正規分布を用いた場合の未調整および各内的検証法の C 統計量のバイアス（係数タイプ 1）

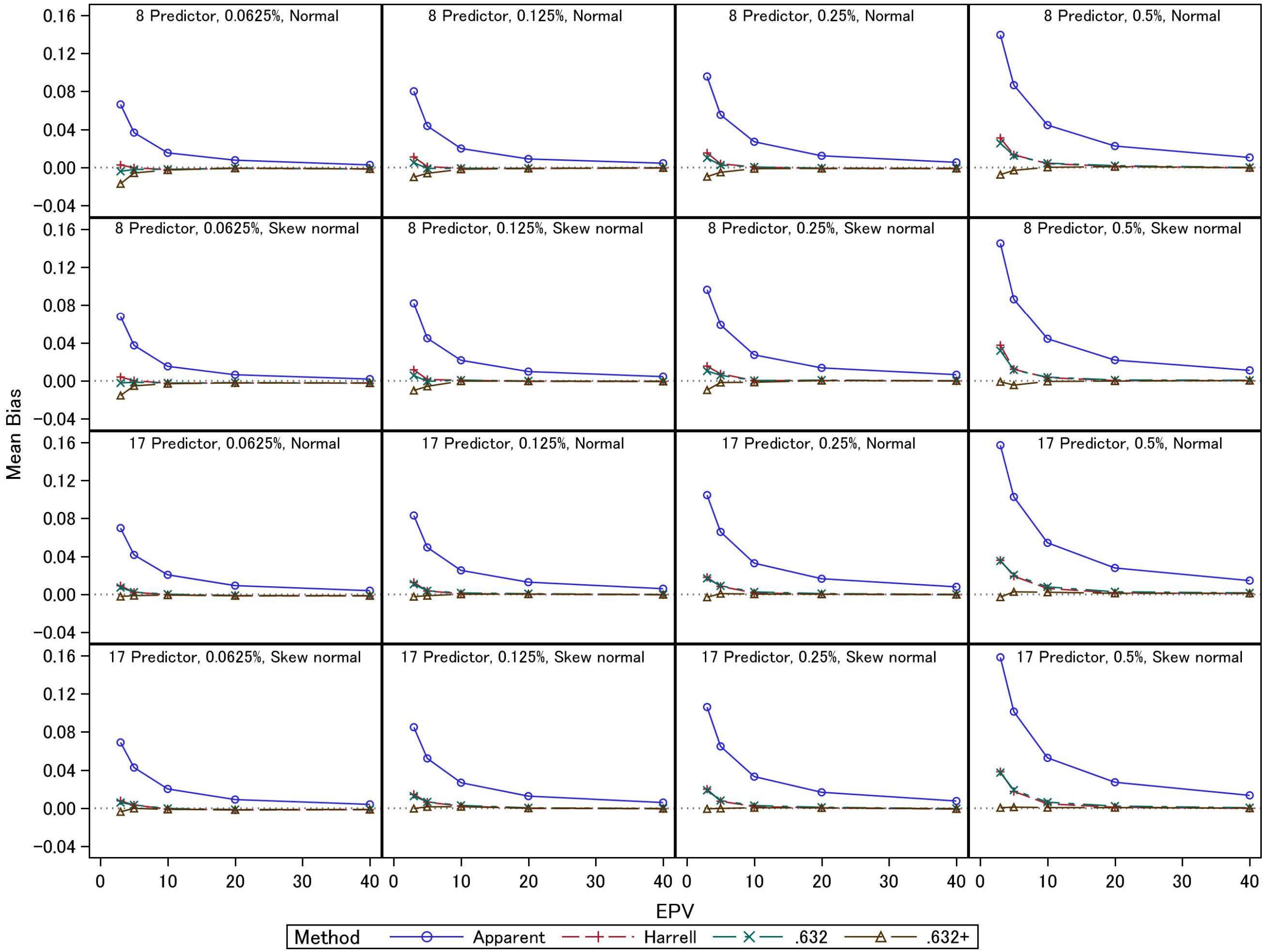

図 3.3.2.4-4 多変量正規および歪正規分布を用いた場合の未調整および各内的検証法の C 統計量のバイアス（係数タイプ 2）

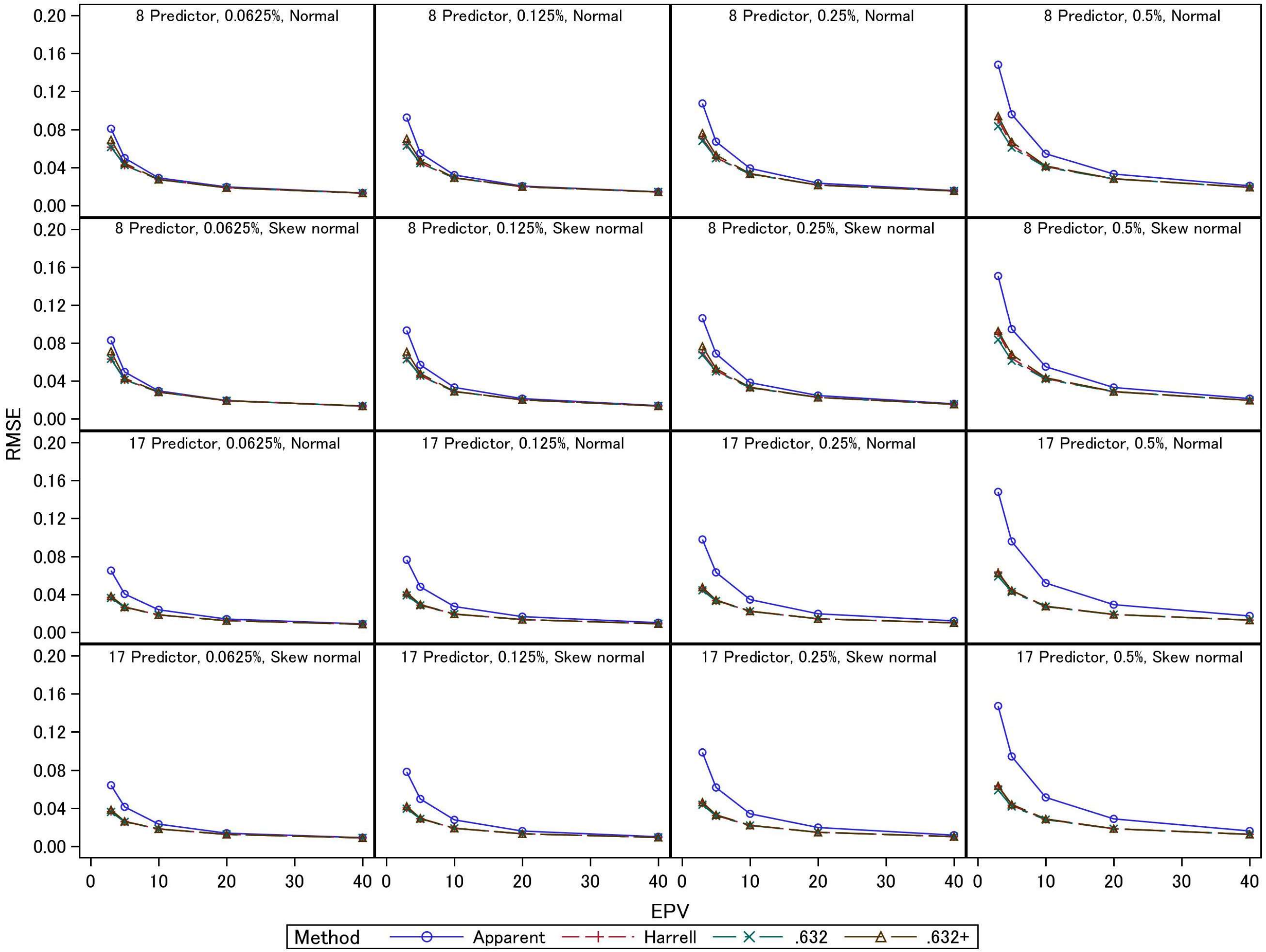

図 3.3.2.4-5 多変量正規および歪正規分布を用いた場合の未調整および各内的検証法の C 統計量の RMSE（係数タイプ 1）

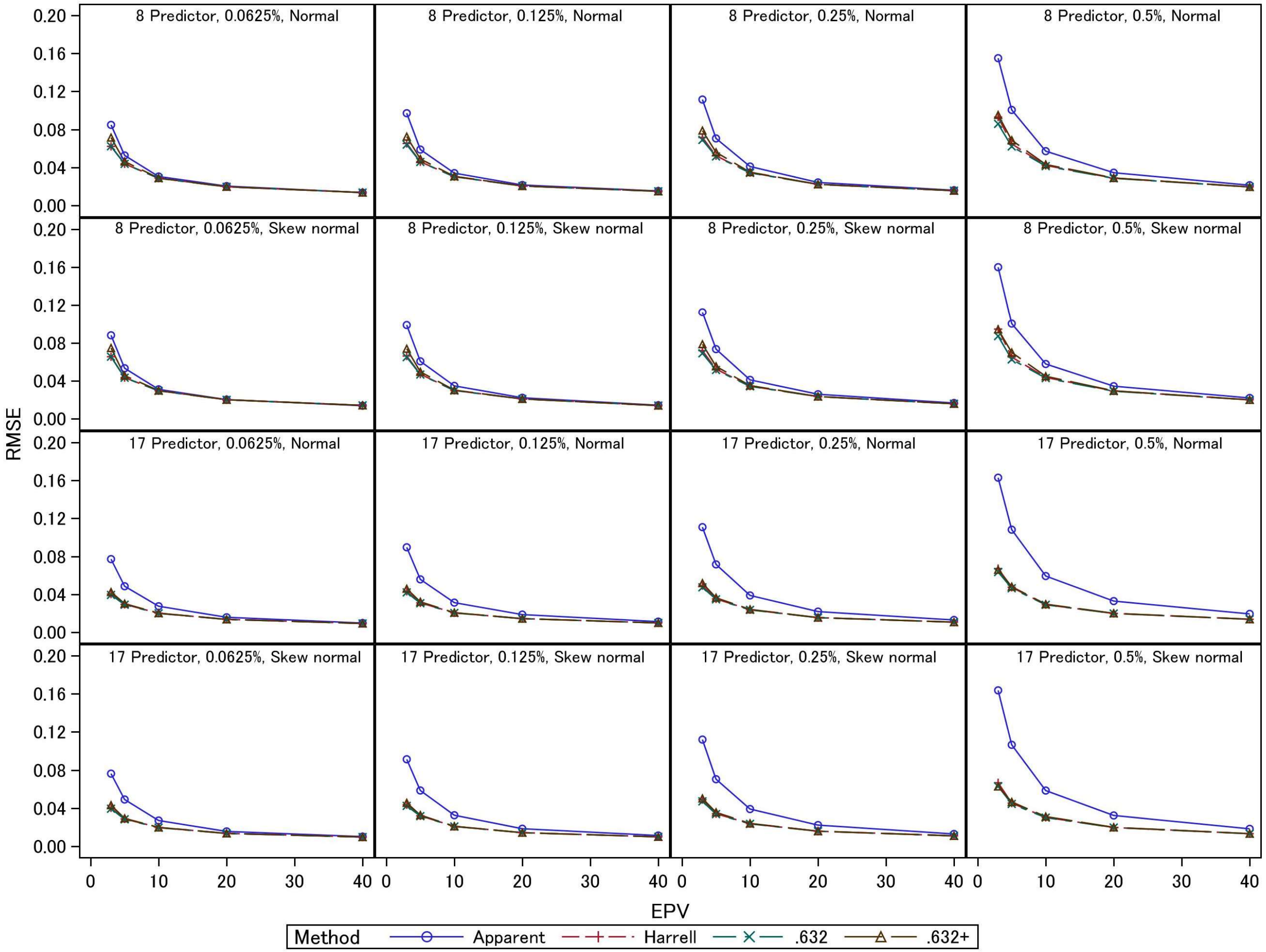

図 3.3.2.4-6 多変量正規および歪正規分布を用いた場合の未調整および各内的検証法の C 統計量の RMSE（係数タイプ 2）

3.4 考察

近年、予測モデルに関する研究報告数は増加傾向にあり、多変量予測モデルの開発は、臨床研究の大きなテーマの 1 つとなっている。多変量予測モデルの開発および報告に関するガイドラインである TRIPOD 声明が公表されたことにより、多変量予測モデルを開発する際に、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によって判別・較正などの予測精度の指標のオプティミスムを補正することが必須となってきている。ブートストラップによる代表的なオプティミスムの調整方法である Harrell 法、Efron の.632 法および.632+法によって得られる推定量は、漸近的に同等であるが、有限標本では異なる特性を持つ可能性がある。このことは、多変量予測モデルを開発する研究の主要な結論に影響する可能性があり、研究の科学的妥当性を保つためには、これらの推定量の有用性に関する確固たるエビデンスが必要とされる。しかしながら、これまでに、これらの推定量の性能を比較・評価した先行研究はわずかしかなく、これらの研究[4, 9]で採用されたモデル構築法およびシミュレーション実験の条件は限定的であったことから、これらの推定量の有用性に関する数値的なエビデンスは十分に得られていなかった。本研究では、広範な実践的条件のもとで、これらの推定量の性能を比較・評価するために、大規模なシミュレーション実験を行った。特に、近年、多変量予測モデルの開発においても普及しつつある正則化法（Ridge 回帰、Lasso 回帰および Elastic-net 回帰）および変数選択のために現在も使用されているステップワイズ法を用いた場合も含めて、これらの推定量の性能について詳細な分析を行った。

従来、多変量予測モデルを開発する際のサンプルサイズの基準として、経験則に基づく EPV ≥ 10 がよく用いられてきた[23]。シミュレーション実験の結果では、EPV ≥ 10 の条件下において、いずれのリサンプリング法に基づく内的検証法も概ね妥当な推定値であった。しかしながら、いくつかの研究において反

例が報告されており[4, 15, 26, 27, 41]、EPV ≥ 10 を絶対的な基準として用いるべきではない。例えば、EPV < 10 において、通常のロジスティック回帰モデルの相対バイアス（回帰係数の真値からの変化率）は 15%以下であり、妥当な推定値が得られる状況があると報告されている[41]。また、近年、Riley et al. [26, 27]は、ロジスティック回帰モデルのオーバーフィッティングを最小限に抑え、かつ全体のイベント発生割合の推定精度を確保するためのサンプルサイズを求める方法を提案しており、その適用事例において、EPV が 10 よりも小さくなる状況もあれば、大きくなる状況もあることを示している。また、ステップワイズ法に対しては、EPV ≥ 10 の基準は十分ではない場合があると報告されている[15]。3.3 節のシミュレーション実験の結果においても、小標本のもとでは、ステップワイズ法の外部の C 統計量は、変数選択を行わない最尤法の外部の C 統計量よりも小さい傾向が認められたことから、ステップワイズ法による変数選択は、実践において推奨されないかも知れない。一方、正則化法（Ridge 回帰、Lasso 回帰および Elastic-net 回帰）は、概ね同等の性能であり、最尤法および Firth 法よりも外部の C 統計量が大きい傾向が認められた。これらのモデル構築法の臨床研究における実用性を評価するために、更なる調査が必要である。

ブートストラップによるオプティミスム補正法に関して、Harrell 法および.632 法は、EPV = 3～5 で顕著な過大評価のバイアスを示した。これらの推定量のバイアスは、イベント発生割合が大きくなると増加する傾向が認められた。3.1 節で述べたように、Harrell 法は、ブートストラップ標本とオリジナル標本のデータのオーバーラップによってオプティミスムを過小評価する可能性があり[8]、小標本のもとでデータのオーバーラップが大きくなることによって、過大評価のバイアスが引き起こされた可能性がある。また、.632 法は、Harrell 法とは異なり、データのオーバーラップによる問題は生じないが、予測

モデルのオーバーフィッティングの程度が強い場合、オプティミスムを補正しきれない問題が知られている[7, 8]。このことから、小標本の条件においてオーバーフィッティングの程度が強くなることによって、過大評価のバイアスが引き起こされた可能性がある。したがって、小標本の場合には、Harrell 法および.632 法の使用には注意が必要である。イベント発生割合 0.5 の条件において正則化法およびステップワイズ法を用いた場合、Harrell 法と比べて、.632 法の過大評価のバイアスが若干大きくなる傾向があった。また、.632+法は、他の条件では過大評価のバイアスを示さなかったが、EPV = 3 において 8 変数モデルのステップワイズ法（$P < 0.05$ の基準）を用いた条件では、過大評価のバイアスを示した。.632 法および.632+法は、未調整の予測精度の指標の推定値と外部標本における予測精度の指標の推定値の重み付き平均によって構成されることから、極端に予測精度が低いモデル（例えば、未調整の C 統計量が 0.5 付近）になったときに、負のバイアスが生じにくい傾向があると考えられた。しかしながら、実践において、そのような極端に予測精度が低いモデルが、最終的な予測モデルとして採用されることはないと考えられることから、この条件で見られた.632+法の過大評価のバイアスは、実践においては大きな問題にならないと考えた。なお、シミュレーション実験全体で、このような極端に予測精度が低いモデルとなってしまうケースが高頻度で起こっていた訳ではない。未調整の C 統計量が 0.6 未満になったケースは、EPV ≥ 10 の条件では認められなかった。また、Ridge 回帰、Lasso 回帰、Elastic-net 回帰およびステップワイズ法において、このようなケースの割合は、EPV = 5 では 0.1-2.1%の範囲（中央値: 0.2%）、EPV = 3 では 0.1-3.6%の範囲（中央値: 0.4%）であった。

シミュレーション実験では、切片のみの予測モデルとなったケースを評価から除外した。切片のみの予測モデルの場合、すべての患者のイベント発生確率の推定値は同じ値となり、カットオフ値によって、すべての患者がイベントも

しくは非イベントと判定される。C 統計量の推定値の定義から、未調整および外部の C 統計量はどちらも 0.5 となり、オプティミスムは 0 となる。これらのことから、切片のみの予測モデルは多変量予測モデルとみなせず、また、オプティミスム補正法の適用対象とならないと考えられる。切片のみの予測モデルが約 10%以上発生した EPV = 3 におけるイベント発生割合 0.5 の 8 変数モデル（Lasso 回帰、Elastic-net 回帰およびステップワイズ法（$P < 0.05$））について、切片のみの予測モデルを除外しなかった場合の結果を示す（表 3.4-1）。なお、これ以外のシナリオでは切片のみの予測モデルの頻度は低いため、切片のみの予測モデルを除外した場合と除外しなかった場合の結果に大きな違いはなかった。切片のみの予測モデルを評価に含めると、上述した予測精度が極端に低いモデルと同様の影響が見られた。切片のみの予測モデルの場合、未調整および外部の C 統計量はすべて 0.5 となるため、C 統計量の平均値は全体的に小さくなった。切片のみの予測モデルはオプティミスムがすべて 0 となるため、未調整の C 統計量の過大評価のバイアスは小さくなり、Harrell 法も同様の傾向を示した。.632 法および.632+法は、上述したように、予測精度が極端に低いモデルでは負のバイアスが生じにくいため、Lasso 回帰および Elastic-net 回帰では過大評価側にバイアスが大きくなる傾向が見られた。一方、ステップワイズ法（$P < 0.05$）では、バイアスは 0 に近かったため、過大評価のバイアスは大きくならなかった。各 C 統計量の標準誤差は、C 統計量の平均値が 0.5 から遠いほど大きくなったが、バイアスの絶対値の減少もあったため、RMSE は必ずしも大きくならなかった。切片のみの予測モデルを除外したことは、シミュレーション実験の全体的な結果を大きく変えるものではないと考えられた。

表 3.4-1 切片のみの予測モデルを除外した場合と除外しなかった場合の結果の比較（EPV = 3、イベント発生割合 0.5、8 変数モデル）

モデル構築法		C 統計量		バイアス		RMSE	
		除外	除外せず	除外	除外せず	除外	除外せず
Lasso 回帰	外部	0.666	0.632				
	未調整	0.796	0.735	0.130	0.103	0.145	0.129
	Harrell	0.709	0.648	0.043	0.016	0.080	0.083
	.632	0.700	0.674	0.034	0.042	0.079	0.083
	.632+	0.667	0.648	0.001	0.016	0.096	0.096
Elastic-net 回帰	外部	0.675	0.653				
	未調整	0.808	0.769	0.133	0.116	0.147	0.138
	Harrell	0.717	0.678	0.042	0.025	0.082	0.084
	.632	0.708	0.691	0.033	0.038	0.078	0.082
	.632+	0.675	0.663	0.000	0.010	0.095	0.096
ステップワイズ法	外部	0.635	0.616				
（P < 0.05）	未調整	0.747	0.712	0.112	0.096	0.129	0.119
	Harrell	0.663	0.626	0.028	0.010	0.078	0.081
	.632	0.676	0.651	0.041	0.034	0.077	0.072
	.632+	0.659	0.637	0.023	0.021	0.077	0.071

.632+法は、予測モデルのオーバーフィッティングの程度が強い場合の.632 法の問題点を克服するために開発された手法であることから、他の 2 つの方法とは異なり、一般的に過大評価のバイアスを示さなかった。しかしながら、.632+法は、イベント発生割合が極端に小さい場合に、若干の過小評価のバイアスを示す傾向があった。これは、イベント発生割合が小さい条件では、少数のイベ

ントがたまたま上手く判別されることによって、オーバーフィッティング率の過大評価が生じることが原因と考えられた。.632+法の過小評価のバイアスを示す傾向は、オーバーフィッティングの程度が強い最尤法で顕著であったが、他のモデル構築法ではそれほど顕著ではなかった。過大評価および過小評価のバイアスは、特にバイアスが量的に大きい場合は、いずれも医療の実践において深刻な問題を生じさせるが、過大評価のバイアスは、実際は予測精度の低いモデルを予測精度が高いと判断し、医療の実践において使用してしまうリスクがあるのに対して、過小評価のバイアスは、実際は予測精度の高いモデルを不採用にしてしまうリスクはあるものの、それは予測モデルを見直すきっかけに過ぎない場合もあり、医療の実践における間違った意思決定に繋がる可能性がある過大評価のバイアスの方が、より大きな問題があると考えられる。

.632+法のバイアスは、一般的に他の 2 つの方法のバイアスよりも相対的に小さかったが、.632+法の RMSE は、他の 2 つの方法の RMSE と同程度か、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。.632+法は、未調整の予測精度の指標の推定値と外部標本における予測精度の指標の推定値の重みを、オーバーフィッティング率で調整することから、小標本のもとでの推定された予測モデルの変動が、.632+推定量のばらつきに寄与していると考えられた。また、.632+法の RMSE は、小標本のもとで正則化法を用いた場合において特に大きい傾向が認められた。正則化法で予測モデルを構築する際、推定値の縮小の程度を調整するチューニングパラメータは、通常 5-fold または 10-fold CV 法によって選択され、本研究のシミュレーション実験では後者を用いた。10-fold CV 法は小標本において不安定であり[42]、これによって予測モデルの変動が大きくなったことが、.632+法の RMSE に影響したと考えられた。また、最新の研究でも、正則化法は、小標本において平均的な予測精度は高いが、チューニングパラメータの推定に大きな不確実性を伴うた

め、信頼できないという報告がある[43, 44]。これらのことから、10-fold CV 法の代わりに、leave-one-out CV 法でチューニングパラメータを選択したところ、Lasso 回帰を用いた場合の.632+法の RMSE は減少した（表 3.4-2）。この結果から、ブートストラップ法に基づく推定量の性能には、チューニングパラメータの選択方法が影響することが示唆された。そのため、小標本の場合、チューニングパラメータの選択方法を慎重に検討する必要がある。なお、10-fold CV 法および leave-one-out CV 法を用いた場合の Lasso 回帰の外部の C 統計量は同程度であり、一般的には 10-fold CV 法の方が推奨されているが、小標本では leave-one-out CV 法を使用することも考えられる。

表 3.4-2 10-fold CV 法および leave-one-out CV 法を用いた際の 8 変数モデルの Lasso 回帰の未調整および各内的検証法の C 統計量の RMSE（EPV = 3）

	係数タイプ	イベント発生割合	未調整	Harrell	.632	.632+
10-fold CV	1	0.0625	0.070	0.047	0.051	0.062
	1	0.125	0.080	0.050	0.052	0.065
	1	0.25	0.102	0.060	0.061	0.076
	1	0.5	0.136	0.077	0.076	0.094
	2	0.0625	0.075	0.048	0.054	0.067
	2	0.125	0.085	0.050	0.054	0.069
	2	0.25	0.106	0.061	0.062	0.079
	2	0.5	0.145	0.080	0.079	0.096
Leave-one-out CV	1	0.0625	0.072	0.054	0.054	0.058
	1	0.125	0.080	0.055	0.055	0.060
	1	0.25	0.096	0.064	0.062	0.067
	1	0.5	0.136	0.084	0.080	0.082
	2	0.0625	0.076	0.054	0.055	0.059
	2	0.125	0.086	0.057	0.057	0.061
	2	0.25	0.101	0.065	0.063	0.067
	2	0.5	0.142	0.086	0.082	0.083

本研究では、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く広範な設定で、大規模なシミュレーション実験を行った。リサンプリング法に基づく内的検証法の性能を比較・評価するために、予測モデルの構築に関わるいくつかの要因を変化させることで、広範な実践的条件を考慮した。本研究の限界は、予測変数の設定が、急性心筋梗塞患者の死亡を評価した GUSTO-I 試験のデータセットのケースのみに基づいており、それらの設定がシミュレーション実験を通して用いられたことである。GUSTO-I 試験は、多変量予測モデルの分野における代表的な 2 値データの事例であり、他のデータセットのケースにおいても本シミュレーション実験の結果は参考にできると考えられるが、他の全てのデータセットのケースでも全く同じ結果になるといった過度の一般化はできない。また、本シミュレーション実験では、GUSTO-I 試験を用いた多くの先行研究[4, 15, 32]で採用されていた 8 変数および 17 変数のモデルのみを考慮した。それ以外のモデルも考慮することは可能であるが、各シナリオで合計 4,000,000 回の反復（2000 回の反復 × 2000 回のブートストラップリサンプリング）が必要であり、シミュレーション実験の計算負荷が非常に大きかったため、他のシナリオおよびデータセットを考慮することはできなかった。他のシナリオおよびデータセットの考慮については、将来の研究における課題である。

また、本研究はリサンプリング法に基づく内的検証法の比較が主な目的であったことから、シミュレーション実験では、すべての候補モデルが真のデータ生成過程含んでいる条件下での評価を優先して行った。しかしながら、実際のデータ解析においてはモデルを誤特定してしまう可能性が考えられる。モデルを誤特定した場合におけるリサンプリング法に基づく内的検証法の性能評価は、今後の重要な研究課題である。

本研究のシミュレーション実験では、広範な実践的条件のもとで詳細な分析を行うために、判別精度の指標として最もよく用いられている C 統計量のみを評価に用いた。内的検証法の評価としては、Brier スコアおよび較正スロープといった他の予測精度の指標も考慮可能である。しかしながら、先行研究において、これらの予測精度の指標は C 統計量と同様の傾向を示していたことから [4]、将来の研究で数値実験による検証が必要ではあるが、他の予測精度の指標についても、本研究のシミュレーション実験の結果と同様の傾向が認められる可能性がある。

結論として、比較的サンプルサイズの大きな条件（EPV ≥ 10）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、いずれにもほとんどバイアスは認められなかった。また、従来の最尤法に加えて、近年、臨床研究においても普及しつつある正則化法を用いた場合や、ステップワイズ法で変数選択を行った場合においても、ブートストラップ法に基づく推定量はいずれも妥当であることが示された。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+法の性能が相対的に優れていた。したがって、一般的には、現在慣例的に用いられている 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。

第 4 章 オプティミスムを補正した信頼区間に関する研究

4.1 オプティミスムを補正法した信頼区間

本節では、提案する位置補正ブートストラップ法および 2 段階ブートストラップ法による信頼区間について説明する。なお、オプティミスム補正法（Harrell 法[1]、.632 法[6]および.632+法[7]）については、3.1.2~3.1.4 で説明した。

4.1.1 位置補正ブートストラップ法

ブートストラップ法の漸近理論[45, 46]に基づいて、予測精度の指標の未調整のブートストラップ信頼区間は、大標本のもとで予測精度の指標のばらつきを適切に評価することができると考えられるが、その位置はオプティミスムによって上方へシフトしていると考えられる。そのため、大標本の場合に位置のバイアスを調整することによって、未調整のブートストラップ信頼区間の問題点について対処できることが期待できる。これらのことから、オプティミスムの推定値によって未調整のブートストラップ信頼区間の位置を調整する位置補正ブートストラップ法を提案した。位置補正ブートストラップ法のアルゴリズムは、以下のとおりである。

1. オリジナル標本に対する未調整の予測精度の指標の推定値を$\hat{\theta}_{app}$、Harrell 法、.632 法および.632+法によるオプティミスムを補正した予測精度の指標の推定値を$\hat{\theta}$とする。
2. $\hat{\theta}$の計算過程で、B 組のブートストラップ標本から$\hat{\theta}_{app}$の標本分布のブートストラップ推定値を得ることができ、これを用いて、$\hat{\theta}_{app}$の未調整のブー

トストラップ信頼区間$(\hat{\theta}_{app,L}, \hat{\theta}_{app,U})$を計算する（95%信頼区間の場合、ブートストラップ分布の 2.5 および 97.5 パーセンタイルである）。

3. オプティミスムの推定値$\hat{\delta} = \hat{\theta}_{app} - \hat{\theta}$を計算する。
4. 未調整のブートストラップ信頼区間を、オプティミスムの推定値によってシフトさせることで、位置補正ブートストラップ信頼区間$(\hat{\theta}_{app,L} - \hat{\delta}, \hat{\theta}_{app,U} - \hat{\delta})$を得る。

位置補正ブートストラップ法の利点は、比較的にシンプルなアルゴリズムで求めることができ、また、オプティミスム補正法の計算過程において、信頼区間の算出に必要な未調整のブートストラップ信頼区間$(\hat{\theta}_{app,L}, \hat{\theta}_{app,U})$およびオプティミスムの推定値$\hat{\delta}$が得られるため、計算負荷が小さいことである。未調整のブートストラップ信頼区間の位置を調整するのは簡単な方法ではあるが、未調整のブートストラップ信頼区間は漸近的に妥当な信頼区間となり[45, 46]、オプティミスムの推定値も大標本になるにつれて 0 に収束することから、位置補正ブートストラップ信頼区間は、大標本理論によって正当化される。しかしながら、未調整のブートストラップ信頼区間では$\hat{\theta}_{app}$のばらつきのみが考慮されているが、オプティミスムを補正した予測精度の指標の推定値$\hat{\theta}$は、オプティミスムの推定値$\hat{\delta}$のばらつき、$\hat{\theta}_{app}$と$\hat{\delta}$の相関によって、$\hat{\theta}_{app}$よりも大きなばらつきを持っている。そのため、位置補正ブートストラップ法は、$\hat{\theta}$のばらつきを過小評価しており、大標本ではない場合においては、被覆確率を名義水準に維持できない可能性がある。

4.1.2 2 段階ブートストラップ法

被覆確率を名義水準に維持するためには、オプティミスムの推定値$\hat{\delta}$のばらつき、および$\hat{\theta}_{app}$と$\hat{\delta}$の相関を適切に考慮しなければならないが、これらの構

成要素間の相関は非常に複雑であり、解析的に評価することは困難である。したがって、これらの変動要因を同時に考慮するには数値的なアプローチが有効であり、オプティミスムを補正した推定量のブートストラップ分布を得る 2 段階ブートストラップ法を提案した。2 段階ブートストラップ法のアルゴリズムは、以下のとおりである。

1. オリジナル標本からのリサンプリングによって、1 段階目の B 組のブートストラップ標本を生成する。
2. 1 段階目の各ブートストラップ標本を用いて予測モデルを構築し、1 段階目の各ブートストラップ標本からのリサンプリングによって、2 段階目の C 組（計 B × C 組）のブートストラップ標本を生成する。
3. 2 段階目のブートストラップ標本を用いて、1 段階目の B 組のブートストラップ標本における Harrell 法、.632 法および.632+法によるオプティミスムを補正した予測精度の指標の推定値$\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_B$を求める。
4. 1 段階目の B 組のブートストラップ標本におけるオプティミスムを補正した予測精度の指標の推定値$\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_B$から、ブートストラップ信頼区間を計算する（95%信頼区間の場合、ブートストラップ分布の 2.5 および 97.5 パーセンタイルである）。

2 段階ブートストラップ信頼区間は、上述した相関を含むオプティミスムを補正した推定量自体のばらつきを適切に評価することができる。4.3 節のシミュレーション実験では、2 段階ブートストラップ法は、未調整のブートストラップ法と比較して、一般的に広い信頼区間幅と名義水準に近い被覆確率を示した。また、2 段階ブートストラップ法は、Harrell 法、.632 法および.632+法のブートストラップ信頼区間に相当することから、その漸近的な妥当性は理論的に

保証されている。しかしながら、2 段階ブートストラップ法は、2 段階のブートストラップリサンプリングを行うため計算負荷が大きく、例えば、両方の段階で 2000 回のブートストラップリサンプリングを行った場合、合計 2000 × 2000 = 4,000,000 回の多変量予測モデルを構築するための反復計算が必要であり、かなりの計算時間を要する。なお、ブートストラップ信頼区間の精度を向上させる目的で、本方法と同様に 2 段階のブートストラップリサンプリングを行うダブルブートストラップ法[46]があるが、本方法はあくまでオプティミスムを補正した推定量のブートストラップ信頼区間を求める方法であり、全く別のアルゴリズムである。2 段階ブートストラップ法では、1 段階目のブートストラップリサンプリングは、オプティミスムを補正した予測精度の指標の点推定値を求めるために、2 段階目のブートストラップリサンプリングは、その点推定値のブートストラップ分布を求めるために行われる。2 段階ブートストラップ法は、これによって求められたブートストラップ分布のパーセンタイル（95%信頼区間の場合、2.5 および 97.5 パーセンタイル）から信頼限界を計算するアルゴリズムとなっている。

4.2 実データの解析

本節では、実践における提案法の有用性を示すために、3.2 節でも解析した GUSTO-I 試験 Western データセット[5]に提案法を適用した結果を示す。各変数の取り扱いは 3.2 節と同様であり、心筋梗塞の発症後 30 日の死亡の有無をアウトカム変数とし、8 変数モデルおよび 17 変数モデルを考慮した。本節では、モデル構築法として、最尤法、Ridge 回帰および Lasso 回帰を用いた。未調整の C 統計量およびその 95%信頼区間（DeLong 法[19]および未調整のブートストラップ法）、Harrell 法、.632 法および.632+法によるオプティミスムを補正した C 統計量およびそれらの 95%信頼区間（位置補正ブートストラップ法および 2 段

階ブートストラップ法）を求めた。ブートストラップリサンプリングの回数は、両方の段階で2000回に設定した（2段階ブートストラップ法では、合計 $2000 \times 2000 = 4{,}000{,}000$ 回のリサンプリングを行った）。

8変数モデルの結果を表4.2-1に示した。8変数モデルでは、未調整のC統計量から補正されたオプティミスムの推定値は、最尤法では0.009、Ridge回帰とLasso回帰では0.007-0.008であった。DeLong法と未調整のブートストラップ法による95%信頼区間は、未調整のC統計量の周辺に位置し、オプティミスムの影響を受けていることが示唆された。最尤法では、2つの提案法（位置補正ブートストラップ法および2段階ブートストラップ法）によって、ほぼ同等のオプティミスムが補正された95%信頼区間が得られた。また、Ridge回帰では、位置補正ブートストラップ法による95%信頼区間と比べて、2段階ブートストラップ法による95%信頼区間（補正されたC統計量のブートストラップ分布）が上方に移動していた。Lasso回帰では、2段階ブートストラップ法の下側95%信頼限界は、位置補正ブートストラップ法の下側95%信頼限界と比較して上方に移動した。一方で、2段階ブートストラップ法の上側95%信頼限界は下方に移動し、補正されたC統計量のブートストラップ分布がより狭くなったことを示していた。3つのオプティミスム補正法（Harrellのバイアス補正法、.632および.632+法）の間で、結果に大きな違いは認められなかった。

17変数モデルの結果を表4.2-2に示した。17変数モデルでは、未調整のC統計量から補正されたオプティミスムの推定値は、一般的に8変数モデルよりも大きく、最尤法では0.021-0.022、Ridge回帰とLasso回帰では0.018-0.019となった。DeLong法と未調整のブートストラップ法による95%信頼区間は、8変数モデルと同様に、オプティミスムの影響を受けていることが示唆された。最尤法では、2段階ブートストラップ法による95%信頼区間では、位置補正ブートストラップ法による95%信頼区間と比べて、全体的な結果は同様であったが、

若干広い 95%信頼区間が得られた。Ridge 回帰では、2 段階ブートストラップ法の下側 95%信頼限界の位置が上方に移動し、補正された C 統計量のブートストラップ分布も上方に移動したことから、予測精度が高くなっていることが示唆された。Lasso 回帰では、下側 95%信頼限界は大きく異ならなかったが、2 段階ブートストラップ法の上側 95%信頼限界は下方に移動した。この結果は、補正された C 統計量の標準誤差は小さくなったが、Lasso 回帰の強い縮小によって予測精度は低下したことを示唆している。17 変数モデルでも、3 つのオプティミスム補正法の間で、結果に大きな違いは認められなかった。

表 4.2-1 8 変数モデルの C 統計量および 95%信頼区間

	最尤法	Ridge 回帰	Lasso 回帰
未調整			
DeLong 法	0.819 (0.783, 0.854)	0.819 (0.784, 0.855)	0.819 (0.787, 0.857)
未調整のブートストラップ法	0.819 (0.788, 0.858)	0.819 (0.787, 0.858)	0.819 (0.787, 0.857)
Harrell			
位置補正ブートストラップ法	0.810 (0.779, 0.849)	0.811 (0.779, 0.850)	0.811 (0.779, 0.849)
2 段階ブートストラップ法	0.810 (0.777, 0.850)	0.811 (0.787, 0.857)	0.811 (0.784, 0.839)
.632			
位置補正ブートストラップ法	0.810 (0.779, 0.849)	0.812 (0.780, 0.851)	0.811 (0.779, 0.849)
2 段階ブートストラップ法	0.810 (0.777, 0.850)	0.812 (0.788, 0.857)	0.811 (0.784, 0.840)
.632+			
位置補正ブートストラップ法	0.810 (0.779, 0.849)	0.812 (0.780, 0.851)	0.811 (0.779, 0.849)
2 段階ブートストラップ法	0.810 (0.777, 0.850)	0.812 (0.788, 0.857)	0.811 (0.784, 0.840)

表 4.2-2 17 変数モデルの C 統計量および 95%信頼区間

	最尤法	Ridge 回帰	Lasso 回帰
未調整			
DeLong 法	0.832 (0.796, 0.867)	0.831 (0.795, 0.866)	0.831 (0.795, 0.866)
未調整のブートストラップ法	0.832 (0.803, 0.874)	0.831 (0.804, 0.873)	0.831 (0.804, 0.873)
Harrell			
位置補正ブートストラップ法	0.811 (0.782, 0.853)	0.812 (0.785, 0.854)	0.813 (0.786, 0.855)
2 段階ブートストラップ法	0.811 (0.782, 0.858)	0.812 (0.794, 0.856)	0.813 (0.786, 0.848)
.632			
位置補正ブートストラップ法	0.811 (0.782, 0.853)	0.813 (0.786, 0.855)	0.813 (0.786, 0.855)
2 段階ブートストラップ法	0.811 (0.782, 0.857)	0.813 (0.794, 0.856)	0.813 (0.785, 0.848)
.632+			
位置補正ブートストラップ法	0.810 (0.781, 0.852)	0.812 (0.785, 0.854)	0.813 (0.786, 0.855)
2 段階ブートストラップ法	0.810 (0.781, 0.856)	0.812 (0.793, 0.856)	0.813 (0.785, 0.848)

4.3 シミュレーション実験

4.3.1 シミュレーション実験の方法

提案法の妥当性を確認し、従来の未調整の方法との比較を行うために、シミュレーション実験を行った。第 3 章と同様に、GUSTO-I 試験のデータセットに基づいた設定のもとで、シミュレーションデータを生成した。また、第 3 章と同様に、予測精度に影響する可能性がある要因として、EPV（1、3、5、7、10、20 および 40）、イベント発生割合（0.125 および 0.0625）、候補の予測変数の数（8 変数モデルおよび 17 変数モデル）および予測変数の回帰係数（GUSTO-I 試験 Western データセットに対する最尤推定値（係数タイプ 1）および Lasso 回帰の縮小推定値（係数タイプ 2））を考慮した。これらの要因を組み合わせた合計 56 のシナリオで検討を行った。2 段階ブートストラップ法の計算負荷が大きいことから、イベント発生割合の設定について、GUSTO-I 試験のデータセットにおけるイベント発生割合である 6.2%に近い 2 つのシナリオについて検討した。また、回帰係数の真値に対する仮定は第 3 章と同様であるが、本研究では、4.2 節の実データ解析で Elastic-net 回帰を使用していないことから、係数タイプ 2 では Lasso 回帰の縮小推定値を設定した。

シミュレーションデータの生成方法は、3.3 節（3.3.1.2）と同じである。

本シミュレーション実験でも、予測精度の指標として C 統計量を用いた。Estimand は、独立に生成した 500,000 例の検証データセット対する外部の C 統計量である。シミュレーション実験の反復回数およびブートストラップリサンプリングの回数は、どちらも 1000 回に設定した（2 段階ブートストラップ法では、合計 1000 × 1000 = 1,000,000 回のリサンプリングを行った）。最尤法によって予測モデルを構築し、DeLong 法、未調整のブートストラップ法、Harrell 法、.632 法および.632+法に基づく位置補正ブートストラップ法および 2 段階ブ

ートストラップ法による 8 つの 95%信頼区間の被覆確率および信頼区間幅を評価した。

4.3.2 シミュレーション実験の結果

シミュレーション実験の結果はネステッドループプロット[47]を用いて表示した。各方法による 95%信頼区間の被覆確率の結果を図 4.3.2-1 に、信頼区間幅の結果を図 4.3.2-2 に示した。

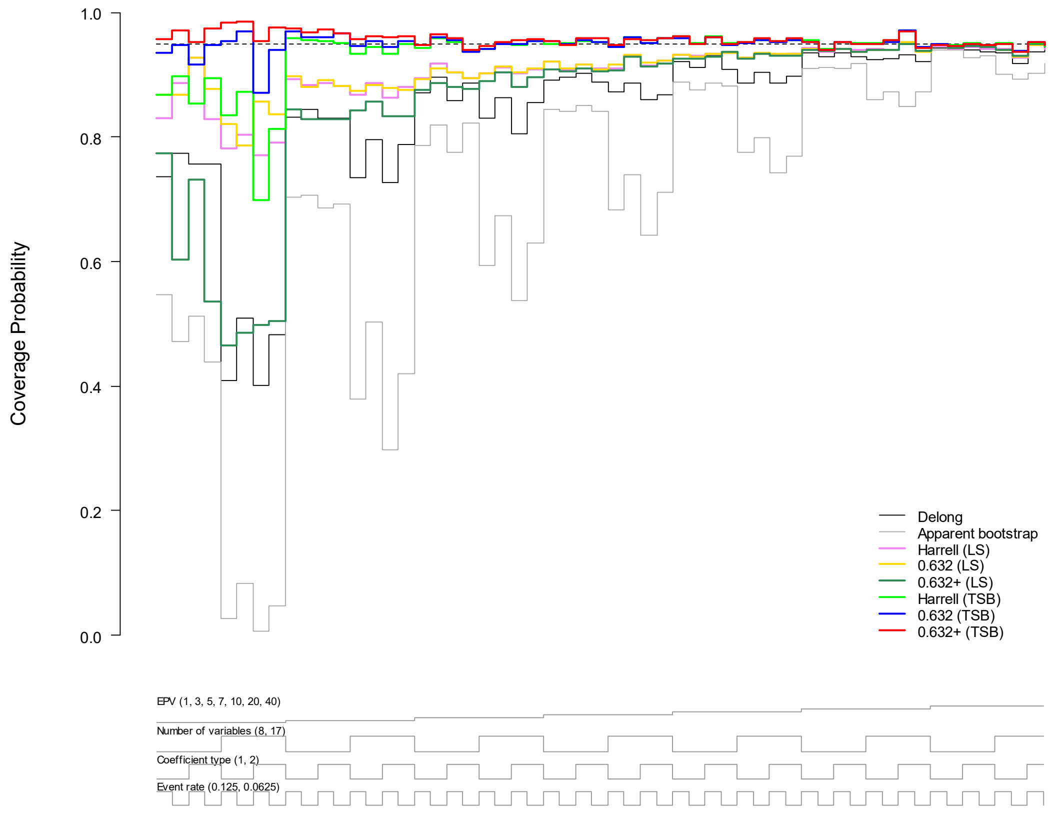

図 4.3.2-1 各方法による 95%信頼区間の被覆確率のネステッドループプロット

ほとんどのシナリオにおいて、未調整のブートストラップ法による信頼区間の被覆確率は、名義水準（95%）を顕著に下回っていた。特に、EPV が小さい

条件や予測変数が多い条件において被覆確率が低くなる傾向が認められた。DeLong 法による信頼区間も同様の傾向を示し、被覆確率は名義水準を下回っていたが、未調整のブートストラップ法と比較して、相対的に大きな被覆確率を示した。これらの結果から、オプティミスムの補正を行っていない未調整の方法は、一般的に不正確な信頼区間を与えることが示された。

提案法の信頼区間は、未調整の信頼区間と比較して、明らかに名義水準に近い被覆確率を示した。位置補正ブートストラップ法による信頼区間は、EPV が比較的に大きい条件（$EPV \geq 10$）では、いずれのオプティミスム補正法でも良好な性能を示し、被覆確率は名義水準に近かった。しかしながら、位置補正ブートストラップ法の被覆確率は一般的に名義水準を若干下回っており、この傾向は EPV が小さくなるほど強く、EPV が極端に小さい条件（$EPV = 1$~3）では、被覆確率が名義水準を顕著に下回っていた。これらの結果は、4.1 節でも述べたように、位置補正ブートストラップ法が、未調整の予測精度の指標のばらつきのみを考慮しており、オプティミスムを補正した推定量全体のばらつきを過小評価する可能性があることが原因と考えられた。しかしながら、位置補正ブートストラップ法は、オプティミスム補正法の計算過程から信頼区間を容易に計算できるにも関わらず、大標本のもとで妥当な信頼区間を与えることが示された。

2 段階ブートストラップ法による信頼区間は、位置補正ブートストラップ法による信頼区間と比較して、より名義水準に近い被覆確率を示した。$EPV = 1$ を除く全てのシナリオにおいて、被覆確率は名義水準付近であり、信頼区間幅は、未調整のブートストラップ信頼区間（および同等の信頼区間幅である位置補正ブートストラップ信頼区間）よりも若干広かった。位置補正ブートストラップ法による信頼区間との被覆確率の差は、EPV が 10 以下の条件において顕著だった。しかしながら、極端に小さい EPV の条件（$EPV = 1$）では、被覆確

率は名義水準を維持できておらず、Harrell 法に基づく信頼区間の被覆確率は顕著に名義水準を下回っていた。また、.632 法に基づく信頼区間の被覆確率も、いくつかの条件において名義水準を下回っていた。一方、.632+法に基づく信頼区間の被覆確率は概ね名義水準であったことから、小標本では.632+法の使用が推奨される。係数タイプ 1 と係数タイプ 2 の結果は同様の傾向であったが、各予測変数のイベント発生のリスクへの寄与が大きい係数タイプ 1 では、被覆確率が若干高くなった。2 段階ブートストラップ法は、ほとんどの条件においてほぼ名義水準の被覆確率を示し、比較的に小標本の場合においても、妥当な信頼区間を与えることが示された。

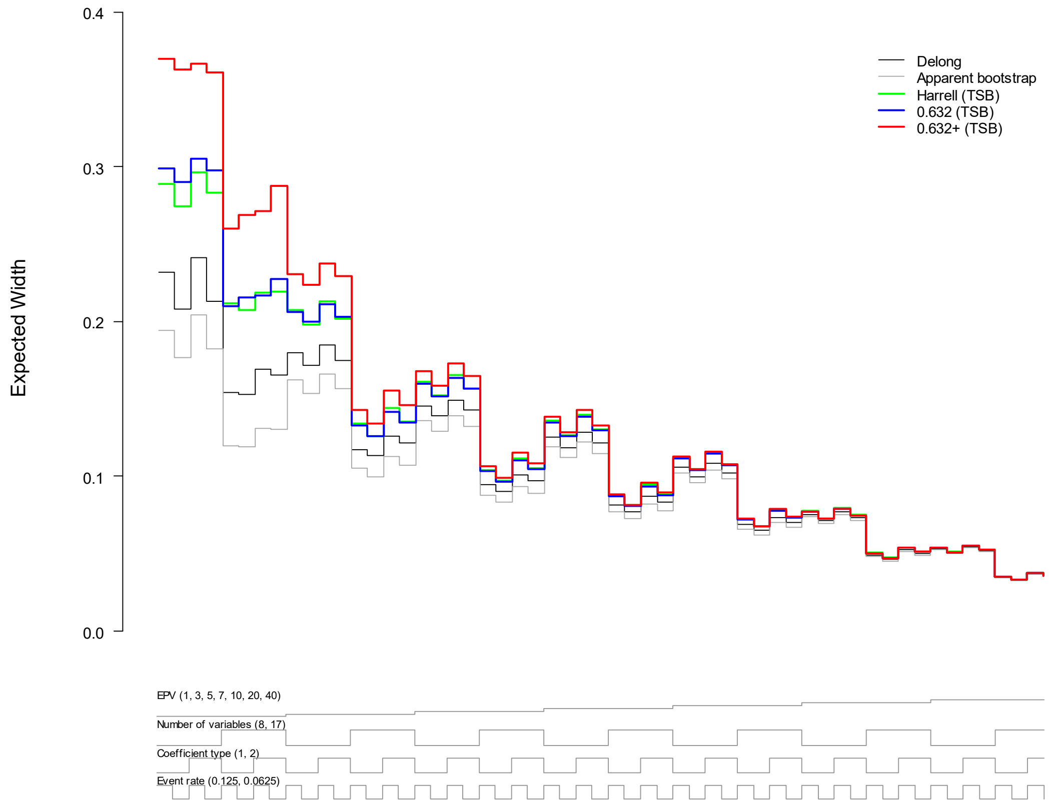

図 4.3.2-2 各方法による 95%信頼区間の信頼区間幅のネステッドループプロット

4.4 考察

近年、多変量予測モデルの開発および報告に関するガイドラインであるTRIPOD声明が公表されたことにより、多変量予測モデルの開発において、判別・較正などの予測精度の指標の内的検証にブートストラップ法を用いることが多くなってきている。現在、ほとんどの臨床研究において、オプティミスムの補正が行われていない予測精度の指標の信頼区間（例えば、C統計量に対するDeLong法）が示されているが、それらは不正確で誤解を招くような結果をもたらす可能性があり、このことは多変量予測モデルを開発する研究の科学的妥当性に影響する可能性がある。シミュレーション実験の結果から、オプティミスムを考慮しない従来の未調整の方法の不正確性が示され、その使用は実践では推奨されず、適切な代替法を用いる必要があると考えられる。

本研究では、この問題を解決するために、2つのオプティミスムを補正した信頼区間を構成する方法を提案した。2つの提案法のうち、オプティミスムを補正した推定量のばらつきを適切に考慮しており、シミュレーション実験の結果からも妥当な信頼区間を与えることが示されている2段階ブートストラップ法の使用がより推奨される。ただし、2段階ブートストラップ法は、2段階のブートストラップリサンプリングが必要であることから計算負荷が大きく、並列計算可能な高性能コンピュータなどを利用しないと実際に適用することは難しいと考えられる。しかしながら、コンピュータの性能は現在も常に向上しており、将来的にはこの問題点は解決されると考えられる。

一定以上の規模のサンプルサイズでは、代替の方法として位置補正ブートストラップ法を使用することができると考えられる。4.1節で述べたように、位置補正ブートストラップ法は、オプティミスムを補正した推定量のばらつきを過小評価する可能性がある。しかしながら、一定以上の規模のサンプルサイズでは、その被覆確率は名義水準に近づき、シミュレーション実験の結果から、

被覆確率は EPV $\geq$ 20 では良好であり、EPV = 10 でも許容範囲であった。また、シミュレーション実験では、位置補正ブートストラップ法の被覆確率は、Delong 法や未調整のブートストラップ法の被覆確率よりも明らかに優れていた。更に、位置補正ブートストラップ信頼区間は、オプティミスム補正法の計算過程において、信頼区間を構成するために必要な値がすべて得られ、追加の計算負荷を必要としないという利点がある。

結論として、多変量予測モデルの予測精度の評価において、従来の未調整の方法は、オプティミスムによって信頼区間の位置にズレが生じており、現実的な条件下では被覆確率は名義水準を大幅に下回っていることから、実践において使用が推奨されず、適切な代替法を用いるべきである。本研究で提案した位置補正ブートストラップ法および 2 段階ブートストラップ法は、どちらの方法も従来の未調整の方法の性能を上回っており、小標本において位置補正ブートストラップ法の被覆確率が名義水準を下回る点を除いては、妥当な方法であることが示された。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。

第 5 章 まとめ

近年、予測モデルに関する研究報告数は増加傾向にあり、多変量予測モデルの開発は、臨床研究の大きなテーマの 1 つとなっている。多変量予測モデルの開発および報告に関するガイドラインである TRIPOD 声明が公表されたことにより、多変量予測モデルを開発する際に、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によって判別・較正などの予測精度の指標のオプティミスムを補正することが必須となってきている。しかしながら、これまでにブートストラップ法による代表的なオプティミスムの調整方法である Harrell 法、Efron の.632 法および.632+法の有用性を比較・評価した研究は限られており、明確なエビデンスがないまま Harrell 法が慣例的に用いられてきた。また、現在の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おらず、多くの臨床研究において、オプティミスムの補正が行われていない予測精度の指標の信頼区間が報告されていた。

本論文ではこれらの問題に対して、まず 1 つ目の研究課題として、広範な実践的条件のもとで、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、大規模なシミュレーション実験を行った。また、2 つ目の研究課題として、2 つのオプティミスムを補正した推定量に基づく信頼区間の計算方法（位置補正ブートストラップ法および 2 段階ブートストラップ法）を提案した。

オプティミスム補正法の評価に関する研究では、比較的に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、いずれにもほとんどバイアスは認められなかった。しかしながら、小標本のもとでは、正則化法が用いられる場合にばらつきが大きくなる点を除いては、.632+法の性能が相対的に優れていた。したがって、一般的には、現在慣例

的に用いられている Harrell 法よりも、.632+法の使用が推奨されることが明らかになった。

オプティミスムを補正した信頼区間に関する研究では、従来の未調整の方法は、現実的な条件下では被覆確率は名義水準を大幅に下回っており、実践では推奨されない方法であることが示された。提案した位置補正ブートストラップ法および 2 段階ブートストラップ法は、どちらの方法も従来の未調整の方法の性能を上回っており、小標本において位置補正ブートストラップ法の被覆確率が名義水準を下回る点を除いては妥当な信頼区間が得られた。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。

これら 2 つの研究から得られたエビデンスにより、多変量予測モデルを開発する研究における予測精度の評価の科学的妥当性の向上が期待できる。

謝辞

博士課程の研究に関して、主任指導教員を引き受けて下さり、常に適切にご助言を頂き、また丁寧にご指導して下さった統計数理研究所の野間久史 准教授に心より感謝を申し上げます。また、副指導教員を引き受けて下さった統計数理研究所の日野英逸 教授に心より感謝を申し上げます。

学位審査においては、お忙しい中審査を引き受けて頂き、数多くの貴重なご意見を下さった統計数理研究所の間野修平 教授および矢野恵佑 准教授、東京大学の菅澤翔之助 准教授に心より御礼申し上げます。

オプティミスム補正法の評価に関する論文の共著者として、大変貴重なご意見を下さった筑波大学の丸尾和司 准教授および東京理科大学の篠崎智大 先生に深く御礼申し上げます。オプティミスムを補正した信頼区間に関する論文の共著者としてご指導頂きました京都大学の古川壽亮 教授および京都府立医科大学の手良向聡 教授に深く御礼申し上げます。東京理科大学の篠崎智大 先生には、本論文に関しましてもご指導いただき重ねて厚く御礼申し上げます。

勤務先の大塚製薬株式会社の新薬開発本部 小野浩昭 本部長ならびに新薬開発本部 クリニカルサイエンス 2 部 池田純司 部長には、博士課程に進学する機会を与えて頂き、深く感謝申し上げます。また、業務に関して配慮下さり、ご支援頂いた所属部門の皆様に深く御礼申し上げます。

最後に、応援して頂き、支えて頂いた家族に心より感謝致します。

参考文献

1. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.

2. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.

3. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63.

4. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models. J Clin Epidemiol. 2001;54(8):774-81.

5. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd edition. New York: Springer; 2019.

6. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc. 1983;78(382):316-31.

7. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc. 1997;92(438):548-60.

8. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition. New York: Springer; 2009.

9. Mondol M, Rahman MS. A comparison of internal validation methods for validating predictive models for binary data with rare events. J Stat Res. 2018;51:131-44.
10. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27-38.
11. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med. 2002;21(16):2409-19.
12. Lee AH, Silvapulle MJ. Ridge estimation in logistic regression. Commun Stat Simul Comput. 1988;17(4):1231-57.
13. Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B. 1996;58(1):267-88.
14. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B. 2005;67(2):301-20.
15. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19(8):1059-79.
16. The Gusto investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. N Engl J Med. 1993;329(10):673-82.
17. Lee KL, Woodlief LH, Topol EJ, Weaver WD, Betriu A, Col J, et al. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. Circulation. 1995;91(6):1659-68.
18. Meurer WJ, Tolles J. Logistic regression diagnostics: understanding how well a model predicts outcomes. JAMA. 2017;317(10):1068-9.

19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. 1988;44(3):837-45.
20. Gart JJ, Zweifel JR. On the bias of various estimators of the logit and its variance with application to quantal bioassay. Biometrika. 1967;54(1/2):181–7.
21. Jewell NP. Small-sample bias of point estimators of the odds ratio from matched sets. Biometrics. 1984;40(2):421-35.
22. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984;71(1):1-10.
23. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.
24. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res. 2019;28(8):2455-74.
25. van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16(1):163.
26. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2019;38(7):1276-96.
27. Riley RD, Ensor J, Snell KIE, Harrell FE, Jr., Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.

28. Akaike H. Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory. 1973:267-81.

29. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461-4.

30. Ambler G, Brady AR, Royston P. Simplifying a prognostic model: a simulation study based on clinical data. Stat Med. 2002;21(24):3803-22.

31. Rahman MS, Sultana M. Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. BMC Med Res Methodol. 2017;17(1):33.

32. Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of shrinkage techniques in logistic regression analysis: a case study. Stat Neerl. 2001;55(1):76-88.

33. Mueller HS, Cohen LS, Braunwald E, Forman S, Feit F, Ross A, et al. Predictors of early morbidity and mortality after thrombolytic therapy of acute myocardial infarction. Analyses of patient subgroups in the thrombolysis in myocardial infarction (TIMI) trial, phase II. Circulation. 1992;85(4):1254-64.

34. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018.

35. Heinze G, Ploner M. logistf: Firth's bias-reduced logistic regression. R package version 1.23. 2018.

36. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1-22.

37. Dai B, Ding S, Wahba G. Multivariate Bernoulli distribution. Bernoulli. 2013;19(4):1465-83.

38. Barthélemy J, Suesse T. mipfp: An R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. J Stat Softw. 2018;86(Code Snippet 2).

39. Azzalini A, Capitanio A. The Skew-Normal and Related Families. Cambridge: Cambridge University Press; 2014.

40. Azzalini A. sn: The Skew-Normal and Related Distributions such as the Skew-t. R package version 16.2. 2020.
41. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007;165(6):710-8.
42. Tantithamthavorn C, McIntosh S, Hassan AE, Matsumoto K. An empirical comparison of model validation techniques for defect prediction models. IEEE Trans Softw Eng. 2017;43(1):1-18.
43. Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Statistical Methods in Medical Research. 2020;29(11):3166-78.
44. Riley RD, Snell KIE, Martin GP, Whittle R, Archer L, Sperrin M, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. J Clin Epidemiol. 2021;132:88-96.
45. Efron B, Tibshirani R. An Introduction to the Bootstrap. New York: CRC Press; 1994.
46. Davison AC, Hinkley DV. Bootstrap Methods and their Application. Cambridge: Cambridge University Press; 1997.
47. Rücker G, Schwarzer G. Presenting simulation results in a nested loop plot. BMC Med Res Methodol. 2014;14:129.