臨床予測モデルにおける Calibration を最適化する解析手法及び変数選択法の研究

On the variable selection methods that optimize calibration in clinical prediction models

都市人間環境学専攻　仕子　優樹

Yuki SHIKO

1. 研究の背景及び目的

個人の持つリスク因子から将来の疾患の発症率やそれによる死亡率を推定するリスク予測モデルは臨床において重要であり、患者、医師または他の医療提供者に、特定の疾患の発症確率を伝え意思決定を支援することに繋がる。臨床予測モデルを作成する際は罹患の有無のような 2 値のアウトカムは、Logistic 回帰、Time to event のデータに対しては Cox 回帰が多く用いられる。モデルで使用する変数は臨床的な側面または統計的手法(変数選択のステップを経て)により決定されモデルの構築がされ、作成した臨床予測モデルは主に discrimination、calibration の観点から評価がなされる。Calibration とは、モデルによる予測確率と観測確率の一致度を表す概念、discrimination はイベントありとイベントなしの対象者を分別する能力を表す概念である。レビュー論文の報告によれば公開されている 78 の論文のうち discrimination の報告は 21 本(27%)で評価・報告がなされておらず、さらに 53 本(68%)で calibration の評価に関する報告がなされていなかった(1)。しかしながら、意思決定をサポートすることを目的とする場合、臨床予測モデルの calibration 能力は重要である。例えば、心血管疾患の 10 年リスクを予測する 2 つのモデル（QRISK2-2011 モデルと NICE Framingham モデル）の外部妥当性を検証した論文では、モデル間の discrimination 能力に大きな違いはなかったが、従来の閾値である 20%を用いた場合(≥20%を高リスク群に割り振る)、QRISK2-2011 モデルでは、35～74 歳の男性 1000 人あたり 110 人が高リスク者と抽出した一方で、NICE Framingham モデルはほぼ 2 倍である 206 人を高リスク者として抽出していた(2)。つまり、この研究は不十分な calibration モデルを使用すると過剰治療につながることを示している。場合によっては過小治療につながる可能性があることを示唆している。臨床予測モデルで用いられる calibration 評価指標には様々あるが、Hosmer-Lemeshow 統計量が最も用いられている。この方法は集団を任意の個数のグループ分けを行い、リスクグループごとの観測度数と期待度数をもとに乖離度の評価を行う。そのため、Hosmer-Lemeshow 統計量は、分割方法によって統計量が左右されることが指摘されている。一方で、近年 Austin PC らにより提案された calibration 指標である Integrated Calibration Index (ICI)は loess 平滑化法を用いて、イベント発生の観測確率を推定し、観測確率と予測確率の絶対差を予測確率の経験分布で重み付けて積分を行うため、このような問題点を解決できる(3)。ICI は単一の基準がないため、作成したモデルの calibration が良いかの判断には使用できないが、他のモデルとの比較に用いることが可能であり、近年いくつかの論文で用いられ始めたが、症例数が大きい場合での検討が多く、臨床データにおける ICI を最適にするモデル構築方法についての検討は十分ではない。そこで本博士論文では過小治療・過大治療を減らすため、calibration に重点を置いた臨床予測モデルの構築方法について検討した。

2. 本論文の構成

本論文は、５章で構成される。各章の内容と成果の概要は，以下のとおりである。

・第１章では、研究の背景と目的、その概要を整理し、臨床予測モデルの評価指標および評価方法についてまとめた。
・第２章では、特に中小規模における calibration を最適にするための解析手法および変数選択法について２つの臨床データを用いて検討を行った。
・第３章では、特にイベント数に対し候補となる因子が多い場合において、calibration を最適にするための罰則付き回帰手法を検討した。さらに、臨床予測モデルは、少ない予測因子(sparse)で優れた予測パフォーマンスがある場合が望ましく、モデルの予測性能(ICI)と sparse 性(選択された予測因子の数)の両方の測定値を１つの図にして提示することで検討した。
・第４章では、臨床の状況を踏まえ、calibration が特に重要である中リスク集団に重きをおいた calibration 指標を定義し、それを最適にするような予測モデル構築方法を提案した。
・第５章では、総括として、本論文の成果および今後の展望についてまとめた。

3. 臨床予測モデルにおける Calibration を最適にする解析手法及び変数選択法：比較研究(2 章)

2 つの規模の異なる臨床データを用い、データセットの規模に応じた ICI を最適にするモデル及び変数選択法の検討を行った。解析手法として 2 値アウトカムに対し臨床予測モデルで良く用いられる Logistic regression モデル (LR)、Linear Probability モデル (LPM)および A Classification And Regression Tree (CART)の３つを用いた。本研究では６つの変数選択法 (Full モデル、 Backward 法(p 値基準, AIC 基準)、Forward 法、Stepwise 法、Least Absolute Shrinkage and Selection Operator (Lasso))を検討した。各臨床データにおいて、データは 70: 30 にモデル作成用データセット(development dataset)とモデル評価用データセット(validation dataset)に分割した。モデルは development dataset を用いて作成し、作成したモデルの calibration 能力の評価(ICI を用いて)を validation dataset において行った。結果、臨床データ 1 では LPM の Lasso 法、臨床データ 2 では LPM における Backward 法(P 値基準、AIC 基準)、Forward 法、Stepwise 法が ICI を最適にしており、データの規模で最適な変数選択法は異なるがいずれの臨床データにおいても LPM の有用性が示唆された。

表１　各臨床データにおけるモデル間の比較

Varibale selection method	Clinical dataset1 (n=137)	Clinical dataset2 (n=478)
	ICI	ICI
LR		
Full model	0.3660	0.0390
Backward method (AIC)	0.4146	0.0396
Backward method (P-value)	0.2004	0.0315
Forward method (P-value)	0.2651	0.0315
Stepwise method (P-value)	0.1419	0.0315
Lasso	0.0636	0.0429
LPM		
Full model	0.1315	0.0460
Backward method (AIC)	0.1194	**0.0308**
Backward method (P-value)	0.1242	**0.0308**
Forward method (P-value)	0.1117	**0.0308**
Stepwise method (P-value)	0.1242	**0.0308**
Lasso	**0.0563**	0.0516
CART	0.1807	0.0516

4. 第3章 臨床予測モデルにおけるCalibrationを最適にする罰則付き回帰法の比較(High-Dimension setting) (3章)

２つの規模の異なるデータを用い、Event per variable (EPV)に応じたICIを最適にするモデル及び変数選択法の検討を行った。２値アウトカムに対して多く用いられるlogistic回帰を用い、４つの罰則付き回帰を比較した (Lasso、Adaptive Lasso、Adaptive double Lasso、Adaptive Elastic Net)。モデルのcalibrationパフォーマンスを評価するためにICIを計算し、内部検証には100ブートストラップ法を使用した。予測モデルは、sparseで優れた予測パフォーマンスがある場合、良いとされる。sparsity と accuracyの定量化は困難であるためモデルのパフォーマンスとスパース性の両方の測定値を一緒に提示することで議論をした (モデルの能力には ICI 及び C-statistics, sparsity には選択された変数の数を用いた)。結果、Adaptive double Lasso が臨床データ 1 (非常に低い EPV) でcalibrationとdiscriminationの最高のparsimonyを達成したことを示し、臨床データ2 (低いEPV)において他の方法に比べ sparse でかつ同程度の予測性能を示した。この結果は、Adaptive double Lasso が高次元設定において臨床予測モデルを開発するための強力な変数選択方法であり、calibrationとdiscriminationのparsimonyを良くすることを示唆している。

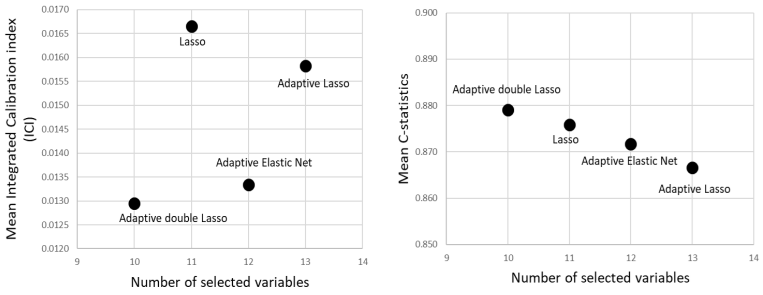

図1　各モデルの性能評価(臨床データ１)

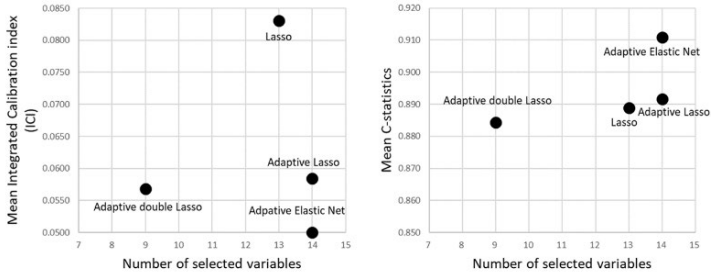

図２　各モデルの性能評価(臨床データ２)

5. 中リスク集団に重点を置いたCalibration指標の開発と最適化 ： Simulation研究(4章)

従来の臨床モデル作成方法では予測確率の全範囲に対して最もあてはまりのよいモデルが選択される。しかしながら、実臨床においては予測確率の全範囲で予測が必ずしも正確である必要はないといえる。たとえば、静脈血栓症 (VTE) などの合併症の発生の予測モデルが臨床診療で使用される場合、患者を低リスク、中リスク、および高リスクに分類するために閾値を設定する。治療方針は、予測モデルから算出される予測確率と閾値に基づいて決定される。具体的には、KearonらはWells scoreとD-ダイマーテスト(101)に基づく肺塞栓症 (PE) の診断テストを提案している。彼らのPEGeDアルゴリズムは次のとおりである；(1) Wells scoreが低く (0〜4.0)、d-ダイマーレベルが1000 ng / mL未満の患者、またはWells scoreが中程度 (4.5〜6.0) で d-ダイマーの患者500 ng / mL未満のレベルでは、

PE の診断テストは行わない。(2) Wells score が高い（6.5 以上）すべての患者が胸部 CT を受ける。このアルゴリズムを使用すると、PE の診断に使用される CT の数が減り、放射線被曝、造影反応、高コスト、時間の消費など、この手順の欠点が軽減される。特にこのアルゴリズムでは中程度のグループにおける Wells score に基づく高い分類精度が必要であると考えられる。中リスク群の患者が高リスクと誤分類された場合、不必要な胸部 CT が実施され、中リスクグループの患者が低リスクに分類されると、D-ダイマーのカットオフが増加し、PE を誤診する可能性が高くなる。つまり、中リスクグループの分類パフォーマンスが重要であるといえる。そこで、本章では、ICI の定義を変更することにより、実際の臨床診療における意思決定を反映する新しい calibration 指標を提案する。さらに、提案された測定を最適化するために、決定木と Logistic モデルの組み合わせたモデルを構築し、提案した指標を最適化することを確認するために、simulation を実行した。結果、決定木と Logistic モデルの組み合わせたモデルがいずれのシナリオにおいても提案した calibration 指標を最適にしていた。

6. 本博士論文の成果

不十分な calibration 性能の臨床予測モデルを現場で用いた場合、過剰治療・過小治療につながるため、臨床予測モデルの calibration 性能は重要である。そこで、calibration 性能を重視する臨床予測モデルを構築するにあたり、本博士論文では臨床予測モデルの calibration 性能を測る指標の 1 つである ICI に着目し、3 つの課題に取り組んだ。2 章・3 章では、臨床予測モデルを構築するにあたり、よくある状況を想定しモデル比較を行っているため、本章の結果は calibration を重点に置いた臨床予測モデルを作成する際の重要な情報となることが期待され、その意義は大きいと考えられる。また、4 章の結果は、臨床予測モデルで用いられている多くの calibration 指標は全確率区間の calibration 一致を測るものであり、臨床を反映した指標は提案されていないことを踏まえると、臨床を反映させた提案指標の提案の意義は大きいと考えられる。さらに今後、複数の診療所、病院、または国からの数千または数百万の患者の電子健康レコードを含むレジストリデータベースが使用可能となる場合、これらの大規模データは外部検証として開発した予測の正確性を確かめるのに最適であると想定される。その場合、予測モデルの作成には対象集団(レジストリデータベース)を意識した構築方法が必要となることが考えられるが、得られたデータを有限母集団とし、その背景の母集団を対象にする Boosting や Penalized 回帰は対象集団を意識していないモデル構成手法であるといえ、適切なモデル作成方法を検討していく必要があるといえる。このような状況では、当てはめたい集団を重視した calibration 指標の活用(ICI の定義を変更しあてはめたいリスク集団での calibration のずれに対する罰則を大きくする等)が考えられ、4 章の検討内容は広く応用可能であると考えられる。

参考文献

[1]Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14(1):40.

[2] Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. BMJ. 2012;344(1):e4181–e4181.

[3] Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine. 2019;38(21):4051–65.