個別の予後や診断に関する多変量予測モデルの透明性ある報告(TRIPOD)のための声明:  解説と詳細

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD):  Explanation and Elaboration

Moons KGM, Altman DG, Reitsma JB, Ioannidis JP a, 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:W1–73.

------------------日本語翻訳検討グループ(敬称略50音順)--------------------

監修：

杉岡隆 (佐賀大学医学部地域医療支援学講座教授)

野口善令 (名古屋第二赤十字病院第一総合内科部長)

福原俊一 (京都大学医学研究科医療疫学分野教授、福島県立医科大学副学長)

翻訳担当：

青木拓也 (京都大学医学研究科医療疫学分野博士課程)

大前憲史 (福島県立医科大学臨床研究イノベーションセンターフェロー、京都大学医学研究科医療疫学分野博士後期課程)

片岡裕貴 (兵庫県立尼崎総合医療センター呼吸器内科、京都大学医学研究科医療疫学分野研究協力員)

小久保雅代 (京都大学医学研究科医療疫学分野博士後期課程)

後藤 禎人 (京都大学医学研究科健康情報学分野医学博士課程)

坂西雄太 (佐賀大学医学部 地域医療支援学講座）

高田俊彦 (福島県立医科大学白河総合診療アカデミー講師、京都大学医学研究科医療疫学分野非常勤講師)

高橋世 (福島県立医科大学臨床研究イノベーションセンターフェロー、京都大学医学研究科医療疫学分野専門職学位課程)

竹島太郎 (自治医科大学地域医療学センター地域医療学部門、京都大学医学研究科医療疫学分野非常勤講師)

辻本康 (京都大学医学研究科医療疫学分野専門職学位課程)

堤悠介 (京都大学医学研究科医療疫学分野専門職学位課程)

長沼透 (福島県立医科大学臨床研究イノベーションセンターフェロー、京都大学医学研究科医療疫学分野博士後期課程)

西脇宏樹 (福島県立医科大学臨床研究イノベーションセンターフェロー)

矢嶋宣幸 (福島県立医科大学臨床研究教育推進部、京都大学医学研究科医療疫学分野 専門職学位課程)

山本舜悟 (神戸大学病院感染症内科)

2016 年 4 月 4 日

目次

内容

[抄録 7](#_Toc447343801)

[本文 7](#_Toc447343802)

[診断および予後予測モデル 8](#_Toc447343803)

[予測モデルの開発、検証、更新 10](#_Toc447343804)

[不完全、不正確な報告 12](#_Toc447343805)

[TRIPOD声明 16](#_Toc447343806)

[TRIPOD声明の作成 17](#_Toc447343807)

[TRIPOD声明: 解説と詳細 18](#_Toc447343808)

[この文書の目的と概略 18](#_Toc447343809)

[使用例 21](#_Toc447343810)

[TRIPOD声明の使用 21](#_Toc447343811)

[TRIPODチェックリスト 22](#_Toc447343812)

[タイトルと抄録 22](#_Toc447343813)

[タイトル 22](#_Toc447343814)

[Item 1. 研究目的が多変量予測モデルの開発 かつ/または検証なのか、対象集団、予測されるアウトカムが分かる 。[D; V] 22](#_Toc447343815)

[抄録 24](#_Toc447343816)

[Item 2. 研究目的、研究デザイン、セッティング、対象者、サンプルサイズ、予測因子、アウトカム、統計解析、結果、結論のサマリーが備わっている。[D; V] 24](#_Toc447343817)

[導入部 26](#_Toc447343818)

[背景と目的 26](#_Toc447343819)

[Item 3a. 医学的文脈（診断なのか予後予測なのかを含む）と多変量予測モデルを開発または検証する意義を既存のモデルへのリファレンスを含めて説明する。[D; V] 26](#_Toc447343820)

[Item 3b. 研究目的を開発なのか、検証なのか、両方なのか明記する。[D; V] 29](#_Toc447343821)

[方法 30](#_Toc447343822)

[データソース 30](#_Toc447343823)

[Item 4a. 研究デザインまたはデータソース（例: ランダム化試験、コホート、レジストリーデータ）、可能であれば開発と検証のデータセットを分けて記述する。[D; V] 30](#_Toc447343824)

[Item 4b. 組み入れ開始日、終了日、可能であればフォローの終了期間を含む重要な研究の日付を明記する。[D; V] 35](#_Toc447343825)

[対象者 36](#_Toc447343826)

[Item 5a. 参加施設の数と場所を含む研究セッティングの重要な要素を明記する（例: プライマリ・ケア、二次ケア、一般人口）。 [D; V] 36](#_Toc447343827)

[Item 5b. 対象者の組み入れ基準を記述する。[ D; V] 40](#_Toc447343828)

[Item 5c関連があれば治療の詳細を記述する。[D; V] 41](#_Toc447343829)

[Item 6a　どうやって、いつ評価されたかを含めて、予測モデルで予測されるアウトカムを明示する。[D; V] 42](#_Toc447343830)

[Item 6b予測されるアウトカムの評価に際して、マスク化を行うために実施した行動について報告する。[D; V] 45](#_Toc447343831)

[予測因子 47](#_Toc447343832)

[Item 7a. どうやって、いつ評価されたかを含めて、多変量予測モデルの開発または検証に使われたすべての予測因子をはっきりと定義する。[D; V] 47](#_Toc447343833)

[Item 7b. 予測因子の評価に際して、マスク化を行うために実施した行為について報告する。 [D: V] 49](#_Toc447343834)

[サンプルサイズ 51](#_Toc447343835)

[Item 8. サンプルサイズがどのように決まったか説明する。[D; V] 51](#_Toc447343836)

[欠測値 54](#_Toc447343837)

[Item 9. 補完方法の詳細を含めて、欠測値がどのように扱われたか記述する（例: コンプリートケース、単一補完法、多重補完法）。[D; V] 54](#_Toc447343838)

[統計解析手法 58](#_Toc447343839)

[Item 10a. 予測因子が解析に際してどのように扱われたか記述する。[D] 58](#_Toc447343840)

[Item 10b. モデルのタイプ、モデルの作成過程（全ての予測因子の選択過程）、内的検証の方法を明記する。[D] 62](#_Toc447343841)

[Item 10c. 検証に際しては、予測結果をどのように計算したか記述する。[V] 70](#_Toc447343842)

[Item 10d. モデルの性能を評価するために使用した全ての指標を明記する、重要であれば複数のモデルを比較する。[D; V] 71](#_Toc447343843)

[Item 10e. 検証の過程で、もしなされているのであれば、モデルのアップデート（例: リキャリブレーション）について記述する。[V] 80](#_Toc447343844)

[リスク群 82](#_Toc447343845)

[Item 11. リスク群が作成されていれば、どのように作成されたか詳細を記述する。[D; V] 82](#_Toc447343846)

[Item 12. 検証に際しては、開発とセッティング、組み入れ基準、アウトカム、予測因子の違いを明らかにする。[V] 84](#_Toc447343847)

[結果 86](#_Toc447343848)

[対象者 86](#_Toc447343849)

[Item 13a. 対象者のアウトカムの有無と人数、可能であればフォローアップのまとめを含めて研究を通じた対象者の流れを記述する。ダイアグラムが手助けになるだろう。［D; V］ 86](#_Toc447343850)

[Item 13b.予測因子とアウトカムについて欠測がある対象者を含めて、対象者の特徴（基本的な人口統計、臨床的な特徴、可能な予測因子）を記述する。[D; V] 89](#_Toc447343851)

[Item 13c　検証に際しては、重要な変数（人口統計、予測因子、アウトカム）の分布を含めて開発のデータとの比較を提示する。[V] 92](#_Toc447343852)

[モデルの開発 95](#_Toc447343853)

[Item 14a. それぞれの解析における対象者とアウトカムの数を明示する。[D] 95](#_Toc447343854)

[Item 14b. もしなされているのであれば、それぞれの予測因子候補とアウトカムの調整抜きの相関を報告する。[D] 97](#_Toc447343855)

[モデルの仕様 99](#_Toc447343856)

[Item 15a.　個人での予測を可能とする予測モデルの全てを提示する（全ての回帰係数と切片や、任意の時点での基準となる生存率など）。[D] 99](#_Toc447343857)

[Item 15b. 予測モデルをどのように使うか説明する。[D] 102](#_Toc447343858)

[モデルの性能 109](#_Toc447343859)

[モデルの更新 113](#_Toc447343860)

[Item17. もしなされているのであれば、モデルのアップデートからの結果（モデルの仕様、モデルの性能）を報告する。[V] 113](#_Toc447343861)

[考察 114](#_Toc447343862)

[限界 114](#_Toc447343863)

[Item 18. 研究の限界（サンプルに代表性がないこと、予測因子に対してイベントが少ないこと、欠測値など）について議論する。[D; V] 114](#_Toc447343864)

[解釈 117](#_Toc447343865)

[Item 19a. 検証を行う場合、結果を考察する際に、開発データでの性能、およびその他の検証データを言及する。[V] 117](#_Toc447343866)

[Item 19b. 目的、限界、似た研究からの結果、他の重要なエビデンスを踏まえて、全ての結果の解釈を提示する。[D; V] 118](#_Toc447343867)

[見込み 119](#_Toc447343868)

[Item 20. モデルの潜在的な臨床応用および今後の研究の見込みについて議論する。 [D; V] 119](#_Toc447343869)

[他の情報 121](#_Toc447343870)

[補足情報 121](#_Toc447343871)

[Item 21. 研究プロトコルやウェブ計算機、データセットといった補足情報の利用可能性について情報を提示する。[D; V] 121](#_Toc447343872)

[資金源 123](#_Toc447343873)

[Item 22.研究における研究資金源および資金提供者の役割について提示する。[D; V] 123](#_Toc447343874)

[結語 124](#_Toc447343875)

[引用文献 126](#_Toc447343876)

# 抄録

TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) 声明は、診断または予後予測モデルの開発、検証、更新に関する報告の改善を目的とした、22のチェックリストを含んでいる。TRIPOD声明は、研究方法によらず、予測モデルに関する報告の透明性を高めることを目的としている。本解説と詳細は、各itemの意味を明確にするための意義{rationale}を示すほか、予測モデルのバイアスのリスクや臨床での有用性を評価し、透明性の高い報告を行うことの重要性について論じている。TRIPOD声明の各チェックリストが詳細に解説されており、優れた報告例も添えられている。本書は、予測モデルに関する研究をデザイン、実施、解析する際に有用な参考文献も提供している。編集過程を支援し、査読者やひいては読者、予測モデルに関するシステマティック・レビューの研究者を援助するため、論文著者にはチェックリストを埋めることが推奨される。TRIPODチェックリストはwww.tripod-statement.orgからダウンロード可能である。

# 本文

医学の分野では、医療提供者は、特定の疾患や状態が存在する（診断の場合）、もしくは将来特定のイベントが発生する（予後の場合）推定確率をもとに、意思決定の共有{Shared Decision Making}を行っている。診断の場合は例えば、追加検査を行う目的で患者を紹介するため、治療開始のため、患者の症状から重篤な疾患を除外するために、特定の疾患が存在する確率を用いている。予後の場合、予測モデルは、一定期間内に特定のアウトカムや健康状態が発生するリスクをもとに、生活様式に関する計画や治療の決定を行うために用いられる (1–3)。このようなリスクの予測は、臨床試験において、対象者のリスクを層別化する際にも用いられる (4–7)。

診断・予後どちらの場合でも、確率予測は通常、個人を対象に測定された複数の予測因子に関する情報に基づいている (1, 2, 8–10)。信頼できる診断確率あるいは予後確率を予測するためには、単一の予測因子からの情報では不十分なことが多い (8, 11)。ほぼ全ての医学領域において、診断や予後の多変量（リスク）予測モデルは、医師や個人に対して確率予測や意思決定を支援するために、開発、検証、更新、実用化されている。

多変量予測モデルは、特定の個人における特定のアウトカムの存在確率（診断）や将来の発生リスク（予後）の予測因子に関する数学的な方程式である (10, 12)。予測モデルは、リスク予測モデル、予後（予測）指標、予後（予測）ルール、リスクスコアなどと呼ばれることもある(9)。

予測因子は、共変量、リスク指標、予後因子、決定因子、検査結果、独立変数などとも呼ばれる。これらは、人口統計属性（例えば年齢や性）から病歴、身体所見、画像検査所見、生理検査所見、血液および尿検査所見、病理所見、疾患のステージや特徴、ゲノミクス・プロテオミクス・トランスクリプトミクス・薬理ゲノミクス・メタボロミクスといった新たに登場しつつある生物学的測定方法の 結果まで広範囲に及ぶ。

# 診断および予後予測モデル

多変量予測モデルは、診断予測モデルと予後予測モデルの2種類に分類される (**Box A**)。診断モデルでは、予測を行う時点における、ある状態や疾患の存在（もしくは不在）を推定するために、複数の予測因子（多くは診断検査の結果）を組み合わせる (**Box B**)。予測因子は、その状態を有していることが疑われる個人から得られ、用いられる。

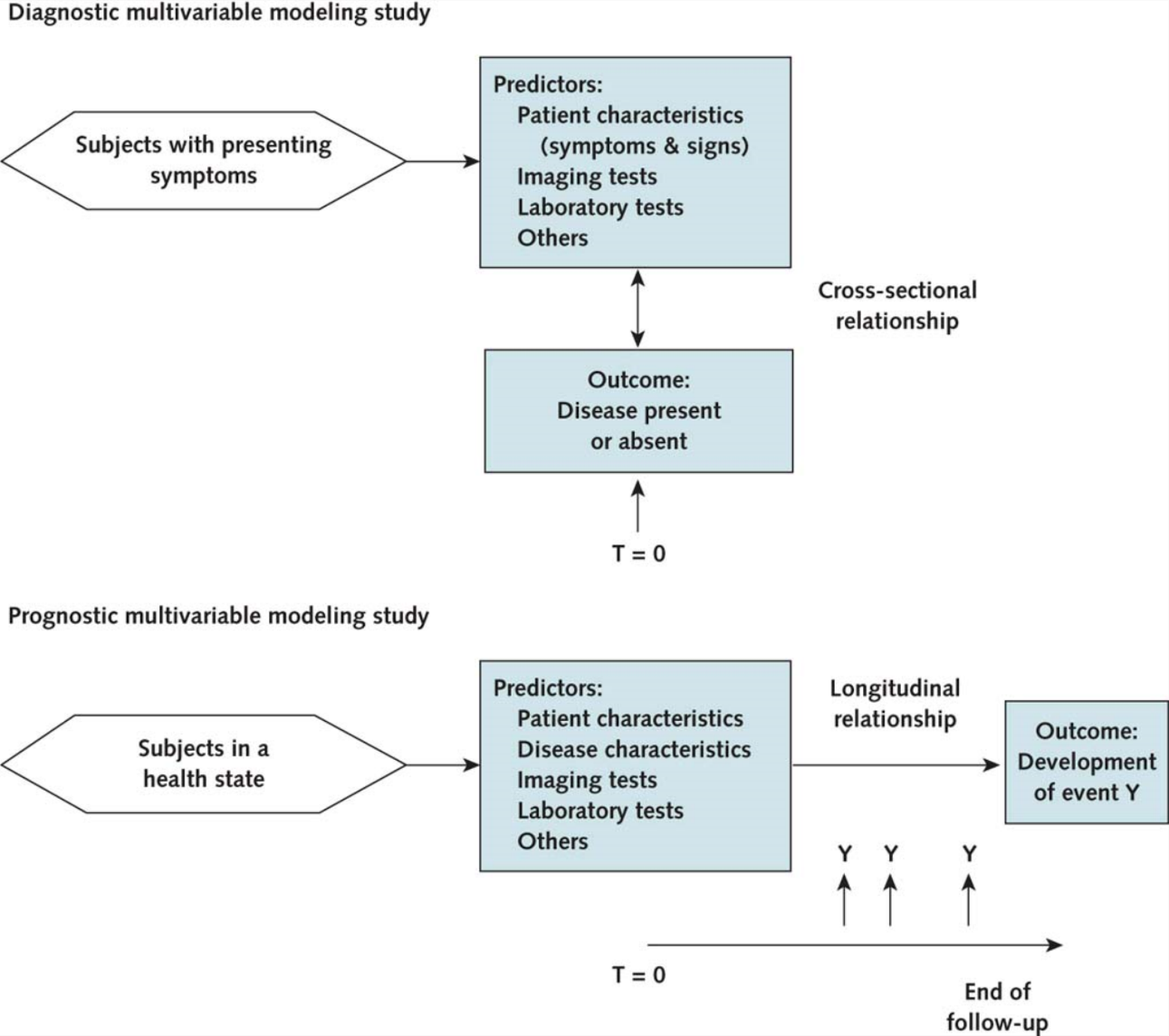
予後モデルでは、将来までのある期間における、特定のアウトカムやイベント（例えば、死亡、疾患の再発、合併症、治療への反応）が発生する可能性を推定するために、複数の予測因子を組み合わせる。この期間は、時間単位（例えば、術後合併症の予測 [13]）から、週や月（例えば、心臓手術後30日以内の死亡の予測 [14]）、年単位（例えば、2型糖尿病の5年発症リスクの予測 [15]）に及ぶ。

予後モデルは、アウトカムが発生するリスクを持つ個人から開発される。その対象は疾病を有するもののこともあれば、健常人のこともある。

例えば、予後モデルには、特定の疾患と診断された後、一定期間における再発や合併症、死亡を予測するモデルが含まれる。しかし予後モデルには、特定の疾患を有さない個人において、一定期間内でのアウトカムの発生を予測するモデルも含まれる。例として、2型糖尿病の発症 (16)や健康な中年の心血管イベント発生 (17)、妊婦の妊娠高血圧症候群の発症 (18)が挙げられる。そのため我々は予後という言葉を、治療の有無にかかわらず特定の疾患を持つ患者の経過予測という狭い定義でなく、リスクを有する個人における将来のアウトカム発生の予測という広い意味で用いる (1)。

診断予測モデルと予後予測モデルの大きな違いは、時間の概念である。診断モデルの研究は通常横断的だが、一方予後モデルの研究は縦断的である。本文書では、診断および予後予測モデルを、ともに「予測モデル」と呼び、それぞれのモデルに特徴的な問題を強調する。

#### Box A 診断および予後予測モデル研究の概略図



診断予測の本質は、個人において現在、予測の瞬間(T=0)における特定のアウトカムや疾患が存在する（あるいは存在しない）確率を推定することである。

予後の場合、予測は個人が一定期間内に特定のイベントやアウトカムを経験するか否かについてである。言い換えれば、診断予測においては関心が横断的関係にあり、一方予後予測においては縦断的関係にある。それでも実際には、実施上の理由によって、診断予測モデルの研究においても、予測因子（診断指標{index test}）の測定と参照基準{reference standard}の測定には、時間的間隔が生じることが多い。理想的には、この間隔は可能な限り短く、この間に何らかの治療的介入が行われないようにすべきである。

# 予測モデルの開発、検証、更新

予測モデル研究は、新たな予測モデルの開発 (10)、モデルの評価（モデルの検証）、モデルの更新 [19–21]、またはこれらの組み合わせを扱う (**Box C, Figure 1**)。

モデル開発研究は、多変量モデルに組み込む予測因子の選定およびその組み合わせにより、予測モデルを導き出すことを目的とする。横断研究（診断的）や短期間の予後的アウトカム（例えば30日間の死亡）の場合にはロジスティック回帰分析が、長期間の予後的アウトカム（例えば10年間のリスク）の場合には Cox回帰分析が、それぞれ頻用される。研究は、予測モデルにある予測因子（例えば新たに発見されたもの） (22) を加えた場合の有用性評価を目的とすることもある。

モデル開発と同一のデータを用いて、モデルの予測能を評価すると（見かけ上の予測能と呼ばれる [**Figure 1**]）、過剰適合{overfitting }（候補因子の数に比しアウトカム発生が少なすぎる場合）や予測因子の選択基準の使用によって、過大評価を もたらす傾向がある (23–25)。そのため、新たな予測モデルを開発する研究は常に、開発したモデルの予測能（例えば較正 {calibration}と識別能 {discrimination}）に含まれる過度の期待{optimism}を定量的に評価し、過剰適合を調整するための内的検証{internal validation}を含んでいる。内的検証は、もとの研究サンプルのみを対象とし、ブートストラップ法や交差検証といった方法が用いられる。内的検証は、モデルの開発に必要な段階である (2)。

予測モデルを開発したのち、モデル開発に用いた集団ではなく、別の集団のデータを用いて予測能を評価することが強く推奨される。外的検証{external validation} (**Box C, Figure 1**) (20,26)は、新しい対象者のデータを必要とし、もとのモデル（発表されたモデルや回帰式）を用いてアウトカムを予測し、実際に観察されたアウトカムとの比較が行われる。外的検証には、同一の研究者が集めた対象者のデータが使われることもあり、一般的には同じ予測因子とアウトカムの定義および測定法が用いられる。後半期間に収集されたサンプル（時間的、限定的検証）や、異なる病院や国で別の研究者がデータを収集した場合（残念ながら稀だが [27]）、異なる定義・測定法が使用されている場合がある（地理的もしくは広範な検証）。対象者が似ていても、セッティングが意図的に異なる（例えば二次ケアのセッティングで開発されたモデルをプライマリ・ケア患者対象で評価する）場合や、対象者が全く異なる（例えば成人を対象に開発されたモデルを小児対象で評価する、致死的イベントの予測を目的に開発されたモデルを非致死的イベントの予測で評価する）場合すらある (19, 20, 26, 28–30)。

外的検証によって評価された予測能が低かった場合（例えば系統的な較正の失敗）、検証用データセットをもとにモデルは変更および調整（例えば再較正や予測因子の追加）される (**Box C**) (2, 20, 21, 31)。

単独のデータセットを無作為分割し、予測モデルの開発および検証に用いる方法は、頻繁に行われている。この方法は外的検証の一種として誤った形で信じられているが、利用可能な全てのデータをモデルの開発に使用しないため、不十分な方法である (23, 32)。開発に利用可能なデータが十分に大きいものであれば、時系列でデータを分割し、一方でモデルの開発を、他方で予測能の評価を行う方法（時間的{temporal}検証）が、最も強力である。単独のデータセットの場合、時間的分割とモデル検証は、内的検証と外的検証の中間と考えることができる。

#### Box B 診断予測モデルと予後予測モデルの類似点および相違点

予測の性質（時期）の違いにも関わらず、診断予測モデルと予後予測モデルには以下の様に多くの類似点がある。

・アウトカムは二値変数であることが多い。疾患の有無（診断）や将来のイベント発生の有無（予後）。

・主要な関心は、患者との情報共有や医療者の意思決定を支援するために、複数の予測因子を用いて、個人に存在または発生するアウトカムの確率を推定することである。

・多変量予測モデルを開発する際には、予測因子の選択、モデルの構築法、連続変数の取り扱い、過剰適合の危険性といった同様の課題が存在する。

・モデルの予測能を評価する方法は同じである。

診断および予後モデル研究において、類似の特徴だが用語が異なるものについて以下に要約する。

|  |  |  |
| --- | --- | --- |
| 診断予測モデル研究 |  | 予後予測モデル研究 |
| 診断テスト、診断指標 | 探索変数、予測因子、共変量(X 変数) | 予後因子、指標 |
| ターゲットとなる疾患/傷害 (あり vs なし)  参照基準と疾患検証 | アウトカム (Y変数) | イベント（未来に発生: あり、なし）  イベントの定義とイベントの測定 |
| 部分検証{partial verification} | 欠測アウトカム | フォローアップからの脱落、  打ち切り |

# 不完全、不正確な報告

予測モデルは、医学文献において徐々に増えており (9, 33, 34)、政策立案者も臨床ガイドラインでその使用を推奨しつつある (35– 40)。ある疾患では、アウトカムや標的集団が同じ競合する予測モデルが多数存在する。例えば頭部外傷後のアウトカムを予測するモデルは100以上 (41)、前立腺癌についても100以上 (42)、乳がんの予後については60以上 (43)、糖尿病診断後の心血管イベントについては45 (44)、2型糖尿病の有病や発症については40以上 (45)、心臓手術後の集中治療室在室日数の延長については20のモデルが存在する (46)。

ほぼ全ての臨床領域にわたって多くの予測モデルが発表されているため、既存の報告に対する批判的吟味やそれらの統合が、読者やケア提供者、政策立案者がどの状況でどのモデルを使用すべきか判断するために必要である。

このような評価は、どのように予測モデルが開発・検証されたかが明確に報告されていなければ不可能である (47, 48)。発表された予測モデルの一般化可能性やバイアスリスクが、適切に評価され (49, 50)、他の研究者が同一のデータから結果を再現することが可能であり、必要であれば同じ結果を得るための手段も示されるべきである(51, 52)。多くのレビューが示しているが、幅広い疾患領域に関する予測モデルの開発・検証について、様々な医学誌で発表された報告の質は、総じて低い (3, 34, 41, 43, 45, 46, 48, 53–95)。例えば、癌に関する文献で発表され、新たに開発された予測モデルのレビューによると、モデル開発におけるあらゆる側面に関して提供された情報は不十分であり、失望すべきほどにその報告の質は低かった(54, 55)。同様の結果は、2型糖尿病の有病・発症に関する予測モデル (45)や、影響力が上位6位以内の総合医学雑誌で発表された予測モデル (34)を対象とした最近のレビューでも認められた。

ランダム化比較試験 (CONSORT [96])、観察研究 (STROBE [97])、腫瘍マーカー研究 (REMARK [98])、分子疫学研究 (STROBE-ME [99])、診断精度に関する研究 (STARD [100])、遺伝的リスクの予測に関する研究 (GRIPS [101]) といった研究報告のためのガイドラインには、予測モデルの開発と検証を含む全てのタイプの研究に関連する項目が 記載されている。予測モデルに最も関係が深いガイドラインは、REMARKとGRIPSの2つである。しかしREMARKチェックリストは、主に予測モデルではなく予後因子に焦点を置いており、GRIPSは、遺伝的リスク因子を用いた予測および数多くの遺伝的変異体の取り扱いに関する特殊な方法論の取り扱いを目的としている。

幅広い研究領域に対応するため、我々はTRIPODガイドライン: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosisを作成した。TRIPODは、診断および予後予測モデルの開発と検証に関して、全ての医学領域とあらゆる種類の予測因子を、分かりやすく網羅している。またTRIPODは、モデル検証研究とその報告の強化に重点を置く。

#### Box C予測モデル研究のタイプ

**開発セット以外の データによる検証を行わない予測モデルの開発研究**

手元にあるデータ（開発セット）を用いて、1つ以上の予後または診断予測モデルを開発することを目的とする。このような研究は、重要な予測因子の同定、多変量解析における予測因子毎の重み付けの調整、個別化された予測に用いるモデルの開発、開発セットでのモデル予測能（識別や較正、分類）の評価を行うことを目的とする。時に開発は、ある予測因子（例えば新たに発見されたもの）の追加による予測能の変化を評価する場合もある。開発研究では、特にデータセットが小さい場合に、過剰適合が起こることがある。そのため開発研究には、ブートスラップ法、ジャックナイフ法、交差検証法といったリサンプリング法を理想的には含むべきである。これらのリサンプリング法を用いることで、開発したモデルの予測能に対する過度の見積もりと、開発に用いたサンプルの母集団から別の対象者を抽出した場合に期待される予測能を定量化することができる (Figure 1参照)。

これらのリサンプリング法は、開発セット以外のデータを使用しないため（手元にあるデータの中で全てを推定する）、モデルの内的検証と呼ばれることもある。内的検証は常に、モデル開発研究の一部である (Figure 1、Box F参照)。

**開発セット以外のデータによる検証を行う予測モデルの開発研究**

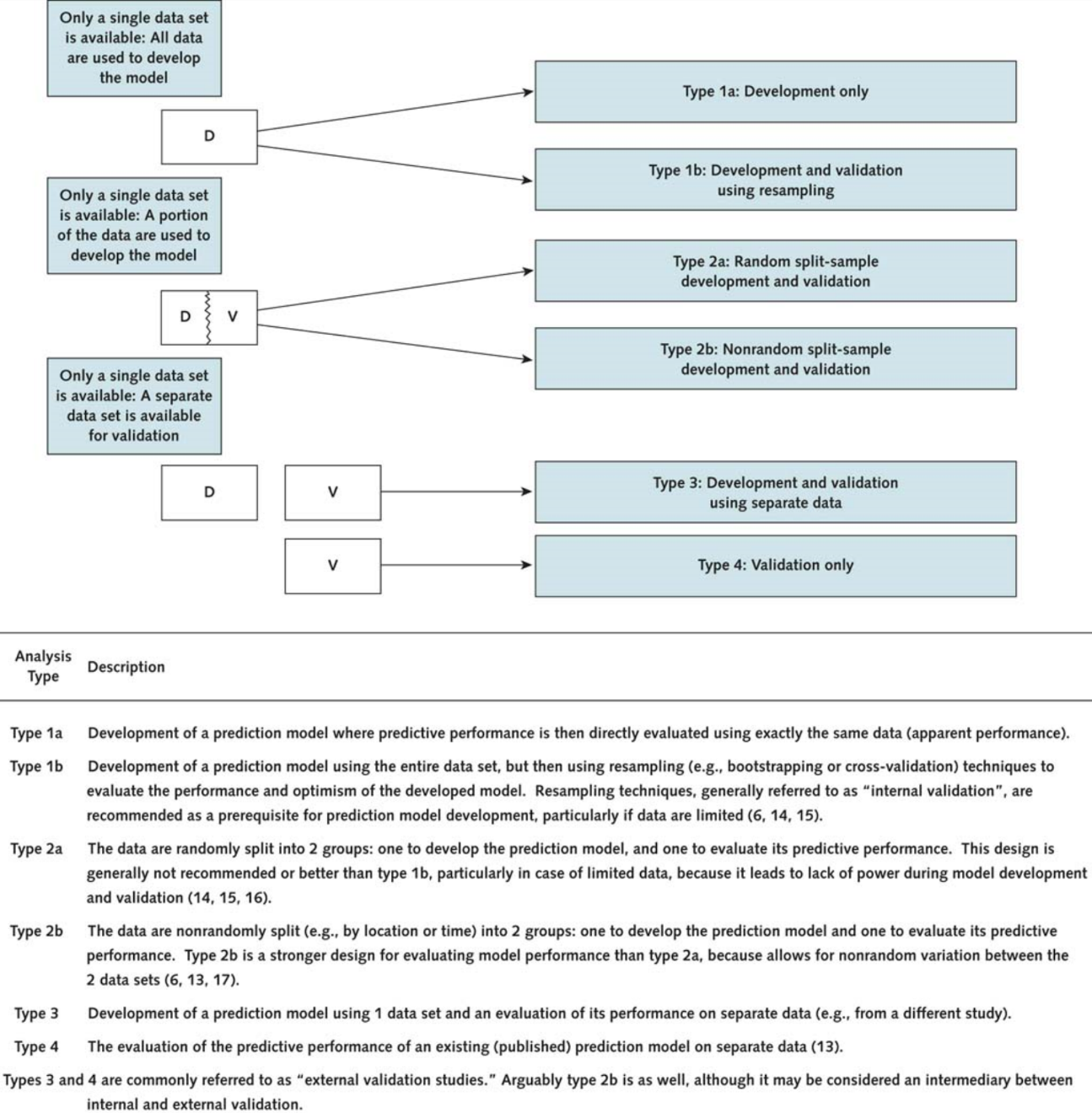
前者のタイプと目的は同じだが、モデル開発に次いで、開発に用いたデータセット以外の対象者データを使用して予測能の評価を行う (Figure 1参照)。これには、同じ研究者が、一般的には同じ予測因子・アウトカムの定義で収集した対象者データが使用されるが、開発後に（時間的あるいは限定的検証と呼ばれる）、異なる病院や国で別の研究者が収集した場合には、異なる定義や測定法が使用されていることがある（地理的もしくは広範な検証）。対象者が似ていても、セッティングが意図的に異なる（例えば二次ケアのセッティングで開発されたモデルをプライマリ・ケア患者対象で評価する）場合や、対象者が全く異なる（例えば成人を対象に開発されたモデルを小児対象で評価する、致死的イベントの予測を目的に開発されたモデルを非致死的イベントの予測で評価する）場合すらある。単独のデータセットを無作為分割し、予測モデルの開発および検証に用いる方法は、外的検証の一種としてしばしば誤って引き合いに出される。2つのデータセットの違いは偶然のみであるため、外的検証というより不十分な形の内的検証である (Figure 1参照)。

**モデルの更新を伴う（または伴わない）モデル検証**\***研究**

モデル開発に使用されなかった対象者データでの予測能の評価や、1つ以上の既存の予測モデルとの比較を目的とする。モデルの予測能が低かった場合、既存モデルの更新や調整（再較正や新たに発見された予測因子の追加）を行う必要がある。初めに元のモデルの検証をしないのは頻度が少なく望ましくないが、原則この研究は既存モデルの更新のみを扱う (Figure 1参照)。

\* *検証*という言葉は、広く使われているが誤解を招きやすい。なぜならこの言葉は、モデル検証研究の結果が"肯定"（検証結果が良い）か"否定"（検証結果が悪い）であることを示唆してしまうからである。しかし、モデル検証の目的は、開発データセットのリサンプルで得られた対象者データ（内的検証と呼ばれる）や、モデル開発で使用されなかった独立した対象者データ（外的検証と呼ばれる）を用いて、モデルの予測能を評価することである。

#### *Figure 1.* TRIPOD声明でカバーされる予測モデル研究の型

**

D = 開発データ; V = 検証データ.

# TRIPOD声明

予測モデル研究は、１）予後あるいは診断予測因子を見つける研究、２）外的検証を行わない予測モデル開発研究、３）外的検証を行う予測モデル開発研究、４）予測モデルを検証する研究、５）モデルの持つインパクトを測る研究、の5つの大きなカテゴリーに分けることができる (1, 8–10, 19, 20, 28, 33, 102–104)。TRIPODでは、１つ以上の予測モデルを開発あるいは検証する研究の報告について取り扱う (Box C)。これら開発や検証のための研究は、さらにいくつかのタイプに分けることができる(**Figure 1**)。また、既存の予測モデルに改訂や調整が必要か調べるために(22, 105, 106)、ある特定の予測因子を加えることで得られる、増分価値を評価するような研究が増えているが(103)、TRIPODではそのような研究も取り扱う (**Box C, Figure 1**)。

予後因子あるいは 診断予測因子を見つける研究やモデルの臨床に与える効果を評価する研究においては、しばしば開発や検証のための研究とは異なる目的やデザイン、報告上重要な点がある。因子を見つける研究では、ある特定の予後的あるいは診断的アウトカムの予測に、独立して寄与する（つまりは関連している）因子を同定しようとする。個別予測のために用いられる最終的な予測モデルを開発することが目的ではない。予測モデルの臨床に与える効果を評価する研究の目的は、そのモデルを用いなかった場合と比較して、モデルの使用が、対象者や医師の意思決定、あるいは直接的に対象者の健康アウトカムにもたらす効果（インパクト）を定量化することである (20, 102, 107)。したがって、モデルのインパクトを測る研究は、開発や検証で用いられるような単一コホートのデザインよりもむしろ比較介入研究のデザインによる。理想的には（クラスター）ランダム化されたものである。この報告ガイドラインで扱うitemの多くは、これら２つの（他とは）異なるタイプの予測研究にも適用できるが、（これらの研究では）おそらく他の報告ガイドラインの方がより役立つであろう。REMARKガイドラインでは（単一の）予測因子研究(98, 108)の報告について明確に述べられている。また、CONSORT声明 (96, 109)はランダム化研究、STROBE声明 (97)は非ランダム化研究の報告に関連のあるガイドラインである。

さらに、TRIPOD 声明では主に、医療分野で最もよく予測されるアウトカムである、2値アウトカム（例えば、疾患の有無）あるいは、時間事象アウトカム（例えば、10年間の心血管疾患発症）に対する予測モデルについて対処する。しかし、予測モデル開発の目的となるアウトカムには、連続変数（例えば、血圧、腫瘍径、血管狭窄率、IQスコア、生活の質、入院期間など）、名義変数（例えば、疾患の有無ではなく鑑別疾患、あるいはウイルス性、細菌性、非感染性などといった感染のタイプなど）、順序変数（例えば、癌のステージ、Glasgow Coma Scale(110)、Rankin scale(111)）も用いられる可能性がある(2, 112)。TRIPOD声明ではほとんどの推奨や報告項目が、このようなアウトカムの予測モデル開発や検証のための研究報告にも同様に適用される。さらに、TRIPOD声明では、医療分野のほとんどの予測モデルの開発や検証、更新の際に使用されるアプローチであることから、回帰モデルを用いて開発された予測モデルを重点的に取り扱う。

しかし、ほとんどの項目は、分類ツリーやニューラルネットワーク、遺伝的プログラミングやランダムフォレスト、あるいはベクターマシン学習テクニックなどのような他の手法を用いて開発、検証、あるいは更新された予測ツールにも同様に適用される。これらの手法と回帰モデルとの主な違いは、予測モデルを導くためのデータ分析手法である。これら非回帰モデルによる手法では、透明性の問題、特に研究の再現性や臨床応用できるかということが特に懸念されるところである。

# TRIPOD声明の作成

我々は、既に公表されている報告ガイドライン作成のためのガイダンス(113)にしたがって、TRIPOD声明の作成を計画し、まとめていくための運営委員会(Dr. Collins, Dr. Altman, Dr. Moons, Dr. Reitsma) を設立した。多変量の予測モデルについての報告や、予測モデルの開発や検証についての方法論的側面、あるいはそれらを評価した多変量の予測モデルに関する報告をまとめたレビュー、または多変量モデルの方法論や報告の仕方について全体的にまとめられたレビューに関して、MEDLINE、EMBASE、PsycINFO、Web of Scienceを用いて系統的に文献を検索した。これらの研究から可能性のある129のチェックリストitemが作られた。そして運営グループが関連のあるitemを統合し、76の候補となるitemのリストを作成した。

さらに、予測モデルに特別な関心のある25人の専門家をe-mailでウェブベースの調査に招聘し、これら76のitemの重要性について順位づけしてもらった。解答が得られたのは27人中24人で、その中には、方法論の専門家やヘルスケアの専門家、雑誌の編集者も含まれていた。（会議の参加者である25人に加えて、Annals of Internal Medcineから2人の統計部門の編集者にも同様の調査を全て行ってもらった。）

24人の専門家（うち22人は既に調査に参加していた）が2011年6月イギリスのオックスフォードで行われた3日間の会議に出席した。統計や疫学の専門家、方法論の専門家やヘルスケアの専門家、そして雑誌の編集者から成る多領域の専門家のグループであった(Appendix)(114)。また、グル－プのうち数人は、他のタイプの臨床研究に対する報告ガイドラインを開発した経験があった。

会議では、調査の結果が発表され、76の候補となるitemについて、順に議論された。それぞれのitemに対して、残すべきか、他のitemと統合すべきか、あるいは省くべきかに関し合意が得られた。会議出席者にはまた、他に追加すべきitemがあるかに関しても質問があった。会議後には何度もの面談を重ねて、運営委員会によるチェックリストの修正が行われ、それが議論を反映しているか参加者に回して確認が行われた。修正の際には、我々の推奨と他のガイドラインとのすり合わせに力が注がれ、可能であれば、itemに対し同じ、あるいは類似した言葉を採用した。

# TRIPOD声明: 解説と詳細

## この文書の目的と概略

TRIPOD声明は、多変量の予測モデルを開発、あるいは検証する研究において、良質な報告を行うために必要と思われる、22のitemからなるチェックリストである(**Table 1**)(114)。それらのitemは、タイトルと抄録に関するもの(item 1, 2)、背景と目的に関するもの(item 3)、方法に関するもの(item 4から12)、結果に関するもの(item 13から17)、考察に関するもの(item 18から20)、そしてその他の情報に関するもの(item 21, 22)となっている。TRIPOD声明は、診断あるいは予後予測モデルの開発だけを報告する研究、開発と外的検証を報告する研究、そして（モデル更新の有無にかかわらず）外的検証だけを報告する研究を取り扱う(Box C)。そのため、予測モデルの開発に関して報告するためだけに関連のあるitem(Dと表示)もあれば(item 10a, 10b, 14, 15)、（外的）検証に関して報告するためだけに適用されるitem(Vと表示)もある(item 10c, 10e, 12, 13c, 17, 19a)。それ以外のitemは全て、あらゆるタイプの予測モデル開発あるいは検証に関する研究に関連するものである。全ての研究のタイプに関連のあるitemはD; Vと表示されている。

TRIPODチェックリスト(**Table 1**)の全22 itemについて、議論と解説が示されている。また、わかりやすくするため、いくつかの複雑で長いitemに関しては、複数の部分に分けて示した。

この解説と詳細に関する既述の最も大切な目的は予測モデル研究の報告を改善するためのフレームワークの概略を示すことである。

しかしながら、多くの予測モデルにかかわる研究は、方法論的に弱い。そのため本稿において、我々は報告の有無に関わらず、良い研究の質（およびあまり良くない研究の限界）についてもまとめている。

#### Table 1 診断または予後に関する多変量予測モデルを開発または検証する研究の報告に際して含まれるチェックリストアイテム

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **セクション／トピック** | **アイテム** |  | **チェックリストアイテム** | **ページ** |
| **タイトルと抄録** | | | | |
| タイトル | 1 | D;V | 研究目的が多変量予測モデルの開発 かつ/または検証なのか、対象集団、予測されるアウトカムが分かる。 |  |
| 抄録 | 2 | D;V | 研究目的、研究デザイン、セッティング、対象者、サンプルサイズ、予測因子、アウトカム、統計解析、結果、結論のサマリーが備わっている。 |  |
| **イントロダクション** | | | | |
| 背景と目的 | 3a | D;V | 医学的文脈（診断なのか予後予測なのかを含む）と多変量予測モデルを開発または検証する意義を既存のモデルへのリファレンスを含めて説明する。 |  |
| 3b | D;V | 研究目的を開発なのか、検証なのか、両方なのか明記する。 |  |
| **方法** | | | | |
| データソース | 4a | D;V | 研究デザインまたはデータソース（例：ランダム化試験、コホート、レジストリーデータ）、可能であれば開発と検証のデータセットを分けて記述する。 |  |
| 4b | D;V | 組み入れ開始日、終了日、可能であればフォローの終了期間を含む重要な研究の日付を明記する。 |  |
| 対象者 | 5a | D;V | 参加施設の数と場所を含む研究セッティングの重要な要素を明記する（例：プライマリ・ケア、二次ケア、一般人口）。 |  |
| 5b | D;V | 対象者の組み入れ基準を記述する。 |  |
| 5c | D;V | 関連があれば治療の詳細を記述する。 |  |
| アウトカム | 6a | D;V | どうやって、いつ評価されたかを含めて、予測モデルで予測されるアウトカムを明示する。 |  |
| 6b | D;V | 予測されるアウトカムの評価に際して、マスク化を行うために実施した行動について報告する。 |  |
| 予測因子 | 7a | D;V | どうやって、いつ評価されたかを含めて、多変量予測モデルの開発または検証に使われたすべての予測因子をはっきりと定義する。 |  |
| 7b | D;V | 予測因子の評価に際して、マスク化を行うために実施した行為について報告する。 |  |
| サンプルサイズ | 8 | D;V | サンプルサイズがどのように決まったか説明する。 |  |
| 欠測値 | 9 | D;V | 補完方法の詳細を含めて、欠測値がどのように扱われたか記述する（例：コンプリートケース、単一補完法、多重補完法）。 |  |
| 統計解析手法 | 10a | D | 予測因子が解析に際してどのように扱われたか記述する。 |  |
| 10b | D | モデルのタイプ、モデルの作成過程（全ての予測因子の選択過程）、内的検証の方法を明記する。 |  |
| 10c | V | 検証に際しては、予測結果をどのように計算したか記述する。 |  |
| 10d | D;V | モデルの性能を評価するために使用した全ての指標を明記する、重要であれば複数のモデルを比較する。 |  |
| 10e | V | 検証の過程で、もしなされているのであれば、モデルのアップデート（例：再較正）について記述する。 |  |
| リスク群 | 11 | D;V | リスク群が作成されていれば、どのように作成されたか詳細を記述する。 |  |
| 開発 vs. 検証 | 12 | V | 検証に際しては、開発とセッティング、組み入れ基準、アウトカム、予測因子の違いを明らかにする。 |  |
| **結果** | | | | |
| 対象者 | 13a | D;V | 対象者のアウトカムの有無と人数、可能であればフォローアップのまとめを含めて研究を通じた対象者の流れを記述する。ダイアグラムが手助けになるだろう。 |  |
| 13b | D;V | 予測因子とアウトカムについて欠測がある対象者を含めて、対象者の特徴（基本的な人口統計、臨床的な特徴、可能な予測因子）を記述する。 |  |
| 13c | V | 検証に際しては、重要な変数（人口統計、予測因子、アウトカム）の分布を含めて開発のデータとの比較を提示する。 |  |
| モデルの開発 | 14a | D | それぞれの解析における対象者とアウトカムの数を明示する。 |  |
| 14b | D | もしなされているのであれば、それぞれの予測因子候補とアウトカムの調整抜きの相関を報告する。 |  |
| モデルの仕様 | 15a | D | 個人での予測を可能とする予測モデルの全てを提示する（全ての回帰係数と切片や、任意の時点での基準となる生存率など） |  |
| 15b | D | 予測モデルをどのように使うか説明する。 |  |
| モデルの性能 | 16 | D;V | 予測モデルの性能を示す指標を信頼区間とともに報告する。 |  |
| モデルの更新 | 17 | V | もしなされているのであれば、モデルの更新からの結果（モデルの仕様、モデルの性能）を報告する。 |  |
| **考察** | | | | |
| 限界 | 18 | D;V | 研究の限界（サンプルに代表性がないこと、予測因子に対してイベントが少ないこと、欠測値など）について議論する。 |  |
| 解釈 | 19a | V | 検証を行う場合、結果を考察する際に、開発データでの性能、およびその他の検証データを言及する。 |  |
| 19b | D;V | 目的、限界、似た研究からの結果、他の重要なエビデンスを踏まえて、全ての結果の解釈を提示する。 |  |
| 見込み | 20 | D;V | モデルの潜在的な臨床応用および今後の研究の見込みについて議論する。 |  |
| **他の情報** | | | | |
| 補足情報 | 21 | D;V | 研究プロトコルやウェブ計算機、データセットといった補足情報の利用可能性について情報を提示する。 |  |
| 資金源 | 22 | D;V | 研究における研究資金源および資金提供者の役割について提示する。 |  |

## 使用例

それぞれのitemに、既に発表された、診断予測と予後予測モデルの開発と検証の両方に関する文献を例示している。そして、報告にあたり有用な情報がどのようなものなのか説明している。我々が挙げた例は、その研究の全ての面が良く構成され、報告されているということを表しているわけではなく、また、そこで報告されている方法が、必ずしも予測モデル研究で最も良い手法であるというわけでもない。むしろ、例として挙げられた研究は、著者が方法の段落の中で上手に報告している、ひとつのitemの注目すべき側面を説明するために用いられている。引用符付きの例のいくつかは本文を編集されていて、…で示される所は本文が省略されていたり、[]で示される所は本文が追加されていたり、また引用文献が省略されたり、あるいは略語が正式に書き換えられたり、いくつかの表が簡略化されたりしている。

# TRIPOD声明の使用

研究者は、予測モデル研究のタイプ（開発なのか検証なのか、あるいはその両方なのか）によって、その研究タイプで関連のあるチェックリストitemについて、結果報告の際にどこかで述べるべきである。もし、個別のチェックリストitemについて述べられないようなことがあれば、その情報が不明なのか、あるいは関連がないのか（本当にそうなのであれば）はっきり報告するべきである。報告の中で、多くのitemが順序立てて述べるべきものではあるが、一部ではそうしていない。それは、雑誌のフォーマット規定にもよるかもしれないので、ここでは我々は、構造化した報告のためのフォーマットは規定していない。著者においては、補足として項を設け、要求された項目に関するいくつかの情報を述べるのも便利かもしれない（例えばオンラインの付録など）。

編集過程や査読者、そして究極には読者を助けるため、我々はチェックリストを追加ファイルとして報告に付けて提出し、そこへは各itemに関する情報がどのページに書かれているか示しておくことを推奨する。TRIPODのチェックリスト報告用テンプレートはwww.tripod-statement.org からダウンロードできる。

TRIPODに関するお知らせや情報はTRIPODのTwitter(@TRIPODStatement)上で広報される。また、The Enhancing the QUAlity andTransparency Of health Research (EQUATOR) Network (www.equator-network.org)がTRIPOD声明の 普及や促進の助けとなる。

# TRIPODチェックリスト

## タイトルと抄録

### タイトル

### Item 1. 研究目的が多変量予測モデルの開発 かつ/または検証なのか、対象集団、予測されるアウトカムが分かる 。[D; V]

**例**

冠動脈疾患を疑わせる患者において冠動脈疾患の可能性を推測するための臨床的スコアの開発と検証(115) [診断; 開発; 検証]

放射線化学療法で治療されている非小細胞肺癌患者の2年生存率を予測するモデルの開発と外的検証(116)[予後; 開発; 検証]

イギリスにおける心血管疾患10年間のリスク予測: QRISK2改訂版の独立した外的検証(117)[予後; 検証]

中年日本人における肝細胞がん10年間のリスク予測モデルの開発: The Japan Public Health Center based Prospective Study Cohort II (118) [予後; 開発]

**追加情報のある例**

運動負荷試験前後での冠動脈疾患を有する確率の増分を推測するためのロジスティック回帰分析から導かれたアルゴリズムの開発と検証(119)[診断; 開発; 検証]

**良く知られたモデルの例**

Framingham冠動脈性心疾患予測スコアの検証: 多様民族グループにおける調査結果(120)[予後; 検証]

イギリス南部におけるSAPS II、 APACHE IIとAPACHE III予測モデルの外的検証: 多施設研究(121)[予後; 検証]

**解説**

潜在的な読者や多変量予測モデル研究のシステマティック・レビュー施行者が検索しやすくするため、タイトルは冗長にならず、できるだけ詳細なものにすることが大切である。著者は、理想的には次の4つの大きな要素をタイトルで述べるとよいかもしれない:

モデル研究のタイプ（開発なのか、検証なのか、あるいはその両方なのか）

臨床的状況（診断か予後か）

対象集団（モデルを当てはめる個人または患者）

モデルで予測されるアウトカム。

予測モデル研究では、モデル開発（内的検証を含む; item 10b）、外的検証、あるいはその両方(**Box C**, **Figure1**)を取り扱う。著者はこれらの用語をタイトルに用いて、研究タイプをはっきりさせるべきである。同様に、その研究が、既存のモデルを更新するもの、あるいはある特定の予測因子を追加した際の増分価値に焦点を当てたものなのであれば、そのように表現することも有用である。なによりも、多くの読者は、ある特定の対象集団、あるいは個人や患者のサブ集団、もしくはこれらの対象におけるある特定のアウトカムに関心があるので、これらが分かるような単語をタイトルに明記することも有用である。

上に例に挙げた、既に発表されている文献例が示すように、タイトルを冗長にすることなく、論文のタイトルでこれらの問題に対処することは可能である。外的検証を含む研究では、それが唯一の目的であろうと予測モデルの開発とともに行われるものであろうと、タイトルにそれもはっきりと示すべきである。診断的か予後的かに関しては、しばしばはっきりと述べられていないことがあるが、対象集団やアウトカムに関する記述の中に暗に含まれている。例えば、タイトルに「冠動脈疾患を疑う男女において…」とあれば、これは診断的研究であることをはっきりと示していると言える(115)。いくつかの予測モデルはその名前を聞くだけでわかるほどよく知られているため、その後の検証研究において対象集団やアウトカムが示されていない。しかし、その研究が、よく知られたモデルを異なる設定で検証したり、異なるアウトカムを予測したりするためのものなのであれば、それはタイトルではっきり示されなければならない。

時に、予測因子のタイプ（例えば、患者の既往歴、あるいは身体診察から得られる予測因子など）、予測のタイミング（例えば、術前の患者の特徴を用いて術後のアウトカムを予測する）、アウトカムのタイミング（例えば、心血管疾患の10年間のリスク）なども、過度にタイトルを長くしてしまうことがなければ、研究の性質をより明確にさせるために、タイトルに含めることも可能である。

最近の、78の外的検証研究のレビューでは、27%(78のうち21)にしか論文のタイトルに、「検証」や「妥当性」といった用語が用いられておらず、また、独立した評価者が検証したことをタイトルではっきり述べているものは、1つしかなかった(122)。

## 抄録

### Item 2. 研究目的、研究デザイン、セッティング、対象者、サンプルサイズ、予測因子、アウトカム、統計解析、結果、結論のサマリーが備わっている。[D; V]

**例**

目的: 外傷性出血のある患者における早期死亡の予測モデルの開発と検証

デザイン: 外傷患者の大規模国際的コホートにおける多変量ロジスティック回帰分析

設定: 40の高所得、中間所得、低所得国における274病院

対象者:

予測モデル開発: The Clinic Randomisation of an Antifibrinolytic in Significant Haemorrhage (CRASH 2) trialにおける受傷後8時間以内の重症出血がある、あるいはそのリスクのある20127人の外傷性患者。

外的検証: イギリスからの患者が主に含まれるThe Trauma Audit and Research Network (TARN)から選択された14220人の外傷患者を対象

アウトカム: 受傷後4週間以内の病院死亡

結果: CRASH 2 trialでは3076人(15%)が死亡し、TARNデータセットでは1765人(12%)が死亡した。Glasgow coma score、年齢、収縮期血圧が最も強力な死亡の予測因子であった。最終モデルのその他の予測因子は、地域（低所得国か中間所得国か高所得国か）、心拍数、受傷からの時間、そして受傷のタイプであった。識別能と較正は、CRASH 2でもTARNでもc統計量で0.80を上回り、条件を満たしていた。治療時点での死亡率が手軽に分かるように、簡単なチャートを作成した。また、より詳細なリスク評価を行うための、ウェブ版計算器が利用可能である。 (http: //crash2.lshtm.ac.uk)

結論: この予後モデルは、外傷性出血患者の死亡を妥当性を持って予測するために使用可能であり、トリアージの助けとなって、診断から救命行為（画像検査や手術、トラネキサム酸投与など）までの時間を潜在的に短縮し得る。年齢は重要な予後因子であり、外傷患者の年齢が高い、高所得国では特に重要である(123)。[予後; 開発]

目的: 急性の足関節外傷において、Ｘ線検査の施行を効率的に行うための既存の臨床決定ルールの検証と改良

デザイン: 2段階で行われる前向き調査: オリジナル版ルールの検証と改良（第1段階）と改良版の検証（第2段階）

セッティング: 2大学病院の救急科

対象者: 簡易{convenience}サンプリングされた、急性足関節外傷のある成人: 第1段階では1130人の適格者のうち1032人、第2段階では530人の適格者のうち453人

主要アウトカムの測定: 救急指導医が各患者に対し、標準的な臨床的変数とオリジナル版ルールによるＸ線検査必要性に対する層別を行い(第1段階)、さらに改良版決定ルールを評価した（第2段階）。標準的な診断基準にもとづく骨折の診断をこの決定ルールで正しく識別できるかどうか、複数の足関節と足のＸ線写真を用いて評価した。オリジナル版決定ルールは単変量及び再帰分割{ recursive partitioning}分析により改良した。

主要結果: 第1段階では、オリジナル版の決定ルールの感度は、121の果部骨折を検出するのには1.0 (95% confidence interval [CI], 0.97 - 1.0)、49の足中央部骨折を検出するのには0.98 (95% CI, 0.88 - 1.0)であった。116人の患者におけるルールを解釈するためのκ係数は、足関節のシリーズで0.56、足のシリーズで0.69であった。20の予測因子の再帰分割により足関節及び足のＸ線検査のシリーズに対する決定ルールの改良を行った。第2段階では、改良版のルールの感度が、50の果部骨折を検出するのには1.0 (95% CI, 0.93 - 1.0)、19の足中央部骨折を検出するのに1.0 (95% CI, 0.83 - 1.0)であった。Ｘ線検査施行の潜在的減少率は、足関節のシリーズで34%、足のシリーズで30%であった。これと同等の決定ルールの結果が「陰性」であれば、骨折の可能性は、足関節のシリーズで 0% (95% CI, 0% - 0.8%)、足のシリーズで0% (95% CI, 0% - 0.4%)と見積もられる。

結論: 改良と検証により、Ottawa足関節ルールは、骨折に対する感度が100%で、信頼性があり、臨床医が足関節の外傷を見た際に、Ｘ線検査のオーダーを1/3減らし得ることが示された。実地試験により、これらのルールを実臨床に導入することが妥当かどうか評価していくことになるだろう (124)。[診断; 検証; 更新]

**解説**

抄録によって読者は研究の方法論と重要さを評価し、所見の要約が分かる。抄録は容易に入手できるので、読者が全文を読むかどうか決める助けとなる。我々は少なくとも研究の目的（理想的には背景あるいは意義についての短い意見により補助される）、セッティング、対象者、サンプルサイズ（とイベント数）、アウトカム、予測因子、統計解析方法、結果（例えばモデルの性能や回帰係数）、結語を含めることを推奨した。個々の雑誌によって求められる形式は様々ではあるが、構造化抄録の方が望ましい。

抄録は、研究がモデル開発、モデル検証、その両者か; 診断モデルか予測モデルか; 標的集団; そして予測されるアウトカムを含めたタイトル（item 1）と同様の特性を示すべきである。モデル開発研究において、もし数が多すぎる場合は全ての候補となる予測因子を明示することは無理かもしれない。この場合は、考えられる総数に言及し、いつそれらが測定されたかを示しつつ（例えば、問診時や診察時）、広いカテゴリーに要約することで充分かもしれない。結果の部分はモデルの予測能の測定に加えて、理想的には最終モデルで含む予測因子を明示するべきである。これは多くの予測因子を含む複雑なモデル、あるいは新しいデータによる既存のモデルへの検証研究には必要ではない。

予測モデル研究の情報に富んだ抄録とタイトルがあれば、研究者は文献検索をする時に、関連のある研究を探し出すことができる。臨床予測モデルを検索する文献検索方式（フィルタ）に関して、いくつか発表されている（125-127）。それらは独立した研究者たちによって最近検証されて、少し修正された。彼らは（他のタイプの予測研究をみつけるにはあまり良くないが）臨床予測モデル研究はあまり見落とさないと結論した（128）。プライマリ・ケアの領域では予測モデルをみつける特異的な検索フィルタも開発されてきている（129）。

## 導入部

## 背景と目的

### Item 3a. 医学的文脈（診断なのか予後予測なのかを含む）と多変量予測モデルを開発または検証する意義を既存のモデルへのリファレンスを含めて説明する。[D; V]

**例**

急性感染性結膜炎を診療するほとんどの総合医は細菌性とウイルス性を区別できないと感じている。実際そのような 患者の80％以上は抗菌薬投与を受ける。これゆえに急性感染性結膜炎の症例において多くの不必要な点眼抗菌薬が処方される．．．．抗菌薬治療から多くの利益をうけるであろう患者を選ぶために総合医は細菌性が原因であることを決定する情報に富む診断ツールを必要としている。総合医がそのようなツールを使うことで抗菌薬処方は減少し、よりターゲットが定まる可能性がある。たいていの総合医は細菌性とその他の原因との間の徴候と症状に基づいて区別をつける。多くは診断の遅れにつながるので、結膜培養のような診断的検査を追加することはめったにない。総合医は実際に細菌性とウイルス性結膜炎を徴候と症状のみに基づいて区別できるのであろうか？・・・最近出版されたシステマティックな文献検索では徴候と症状を要約し、これらの主張に根拠がないことを示した。このペーパーは急性感染性結膜炎の徴候と症状の診断的情報価値についての最初の経験的研究であるようだと示している（130）。[診断; 開発]

耳下腺がん患者の実践的な予後予測システムのための検索では、我々は以前オランダがん研究所からの耳下腺がん患者151名を母集団としたCox 比例ハザード分析に基づく予後指標を作成した。[その]表で．．．治療前予後指標PS1、(手術前に得られる情報を組み合わせたもの)と、治療後の予測指標PS2、(手術検体からの情報を合わせたもの)を示す。それぞれの患者で、この指標は、適切に重み付けした重要な臨床病理学的特徴を合計して、腫瘍再発の推定確率を算出する。これらの指標は作成された母集団および独立した耳下腺がん患者ののオランダ全国データベースにおいて良好な識別能を示した。Justiceらによると、検証の次のレベルは国際的レベルにまでさらに進めることである．．．．この目的のため、国際的データベースはルーヴェンとブリュッセル（ベルギー）とケルン（ドイツ）で治療された患者で構成され、そこでは指標の計算に必要な予後変数が記録され、予測はアウトカムと比較された。このような流れで、我々はさらなる臨床的、統計的検証を達成しよう試みた（131）。[予後; 検証]

リスク予測モデルの改訂と更新に際しては、ルーチンの臨床診療における有用性が劣化していないことを示すため、あるいはその性能がモデル改良によって改善していることを実際に示すために、連続的な評価（検証）をするべきである。

我々は、英国の総合診療記録の大きなデータセットでQRISK2 2011、QRISKの旧バージョンとフラミンガムリスク予測モデル調整版NICEの性能を独立して評価し比較した結果を記述する（117）。[予後; 検証]

**解説**

多変量予測モデルは複数の目的で役立つ、そのため読者はモデルの意義と使用可能性についての明確な記述を必要とする。著者はモデルが使用されるであろう（決断のような）特定の臨床状況を記述するべきである。例えば、診断予測モデルはある患者に、より侵襲的あるいは負担のかかる検査を実施するかどうか決断することに役立つだろう。また予後モデルは将来のアウトカムについてある状況の患者に情報を与え、あるいは次の治療の可能性を判断するのを助けるために使用されるだろう。この医学的状況とその状況におけるモデルの使用は患者（セッティングを含める）の選択と結果が一般化されるであろう対象と、どのタイプの予測因子がこのセッティングで使用可能でそれゆえ考慮されるべきか、について理論的な根拠を与える。アウトカムの選択はモデルの臨床上の重要性を決定している重大な要素である、それゆえに特異的なアウトカムを選択した根拠が示されるべきである。なるべく、アウトカムとフォローアップ期間は患者と臨床上の意思決定にとって関連のあるものであるべきである。

もしより拡大解釈したアウトカムの定義が使われる場合、それによってハイリスクとしてより多くの人を誤分類するリスクが増える問題点がある（132）。同様の問題は診断のセッティングにおいても存在する。もし新しい感度のよいマーカーあるいは高解像度画像検査の異常が新しい疾患の定義になるならば、それは過剰診断や過剰治療につながる（132，133）。医学的状況の記述はリスクの情報に基づいてされる臨床上の決断も表示するべきである。以下診断目的と予後目的両方の多変量予測モデルの異なる使用の例である。

**多変量診断モデルの潜在的な臨床応用:**

より侵襲的あるいはコストのかかる診断検査を指示するかどうか、あるいは患者を二次ケアに紹介するかどうかの決定。

例: 足関節損傷の患者に放射線撮影をいつ指示するかのオタワルール（134，135）。

ある診断したい状態{target condition}が安全に鑑別できるかどうかの決定。

例: 深部静脈血栓症あるいは肺塞栓症を除外する臨床的意思決定ルールとD-dimer（136）。

妊娠中の胎児が21トリソミーを持っている確率について両親への情報

例: 妊娠中のトリプルテスト（137）。

多変量予後モデルの潜在的な臨床応用:

「健康な」個人に対し、心血管疾患の１０年間のリスクに関する情報を与える。この情報は不健康なライフスタイルを変えるために使用されうる。

例: フラミンガムリスクスコア（138），QRISK2（139），SCORE（140）。

ある疾患と診断された患者あるいは特定の外科的手技を受ける患者に対して、事前の予防戦略あるいは治療戦略を決定するための、予後不良や合併症をおこすリスクについての情報。

例: 急性心筋梗塞後の30日死亡率を基準とした血栓溶解療法の適応（141）。

モデルを開発するとき、研究者は理想的には文献レビューに基づき、関連するモデル（例えば、同じあるいは類似した用途、対象、あるいはアウトカム）が既に開発されているのか言及するべきである（47）。外的妥当性研究は既に開発されたモデルの新しい患者に対して存在する性能についての価値ある情報を作成する。

著者は、彼らが妥当性を検証しているモデルが存在することと引用論文を明確に述べるべきであり、またこのモデルの（潜在的な）臨床応用について述べ、言い直すべきである。もし他の競合する予測モデルが存在するなら、著者はなぜ選択したモデルだけを評価したのか示さなければならない。明らかに、同じデータセットでの比較妥当性研究（すなわち複数の競合モデルを評価すること［48］）は追加情報を生むだろう（47，85）。開発されたモデル研究において比較する患者集団、予測因子あるいはアウトカムの慎重な変更は、その意義に沿って強調して表されるべきである（item 12参照）。

外的検証研究の最近のシステマティック・レビューでは、45件中7件（16％）が評価されていた予測モデルを開発するオリジナルの研究を引用していなかった（122）。

### Item 3b. 研究目的を開発なのか、検証なのか、両方なのか明記する。[D; V]

**例**

この研究の目的は乳房の症状がある女性においての臨床予測ルールを開発し検証することである。多くのエビデンスに基づいた紹介へのアプローチ－2週間以内の緊急性のある紹介ルールを含めて－が臨床的診療ガイドラインに載せることができるようになるだろう。[診断; 開発; 検証]

この論文で、我々は早期乳がん患者の長期の無再発生存期間を予測する線形予後モデルの推定と外的検証を新しい英国のデータを元に報告する。モデルの性能は、ノッティンガム予後指標とAdjuvant Onlineと比較し、スコア化アルゴリズムとその活用促進のためにダウンロード可能なプログラムを発表した（143）。[予後; 開発; 検証]

予測モデルについて、新しい患者に対するその予測精度が正式に検証される前に実臨床で適用されるべきはないということが広く受け入れられているにも関わらず、これらの指標について、独立した患者集団での正式で定量的な（外的）検証は未だなされていない。それゆえに我々は、心臓外科手術後の遷延するICU滞在期間（PICULOS）に対する既存の全ての予測モデルを確認するためのシステマティック・レビューを最初に行った。その後にレビューで見つかったモデルの性能を、心臓外科手術患者の大規模な独立コホートで検証した（46）。[予後; 検証]

**解説**

研究目的はその研究で扱われる具体的な目的あるいはリサーチクエスチョンである。目的を明確に具体化することによって、しばしば導入部のおわりに、著者は読者が研究の批判的吟味をする助けとなるような背景情報を与えるであろう。予測モデル研究では、研究目的として、（診断あるいは予後）予測の目的、予測されるアウトカムあるいはアウトカムのタイプ、モデルが使用されるセッティングや意図される集団、考えられている予測因子のタイプを明らかにするべきである。さらに、著者はその報告が新しいモデルの開発あるいは既存のモデルの外的妥当性の検証あるいはその両者に関わるものなのかを記述するべきである。

## 方法

## データソース

### Item 4a. 研究デザインまたはデータソース（例: ランダム化試験、コホート、レジストリーデータ）、可能であれば開発と検証のデータセットを分けて記述する。[D; V]

**例**

1971年から1974年のフラミンガム心臓研究{Framingham Heart Study}調査時点で30歳から74歳の2489名の男性と2856名の女性が含まれたこの報告に使用される一般人口に基づいたサンプルを使用した。対象はオリジナルのフラミンガムコホートの11回目の調査あるいはフラミンガム子孫研究の最初の調査のどちらかに参加したものである。同様の研究プロトコルがそれぞれの研究で使われ、基礎調査で明白な冠動脈疾患がある人は除外された（144）。 [予後; 開発]

多施設国際共同臨床試験（Action in Diabetes and Vascular disease: preterax and diamicron MR controlled evaluation)(ADVANCE) は糖尿病患者の新しい心血管系リスク予測式を作成{derivation}することを可能にする．．．．ADVANCEは２型糖尿病のハイリスク患者11140人を対象とした微小血管あるは大血管系のイベントの発生をアウトカムとした血圧（perindopril indapamide　対　プラセボ）と血糖コントロール（gliclazide MRに基づいた集中的な介入　対　標準ケア）の2×2{factorial}ランダム化比較試験であった．．．DIABHYCAR（The non insulin dependent diabetes, hypertension, microalbuminuria or proteinuria, cardiovascular events, and ramipril study）は1995年から2001年に16ヵ国で実施された2型糖尿病患者に対するラミプリルの臨床試験であった。ランダム化された4912名の対象者のうち、3711名は．．．検証に用いるのに適切であった。DIABHYARの心血管疾患の定義はADVANCEと類似していた．．．．考慮された予測因子は糖尿病診断時の年齢、糖尿病と診断されてからの期間、性別、．．．そしてランダム化した治療（血圧を下げることや血糖コントロールレジメン）（145）。[予後; 開発; 検証]

我々は英国南西部にある5つの救急施設で、急性肘関節損傷にて来院した成人を対象とした多施設前向き検証研究と小児を対象とした観察研究を行った。検査の診断精度は小児では評価されていなかったので、我々はこのグループでの介入研究が正当化されるとは考えなかった（146）。[診断; 検証]

我々は多様なセッティングにおいて慢性閉塞性肺疾患患者の個々の死亡をどのくらい良好に予測するのかを決定するためにADO indexに対するこのように大規模な国際的な検証を行い、必要に応じて更新した。ヨーロッパとアメリカの10の慢性閉塞性肺疾患と集団ベースコホート研究からの研究者が国際的慢性閉塞性肺疾患コホート共同研究ワーキンググループへの協力に同意した（147）。[予後; 検証; 更新]

**解説**

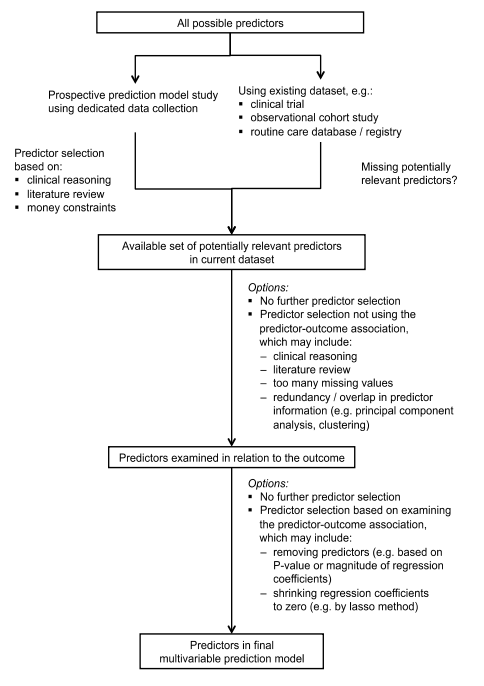
さまざまなデータソースや研究デザイン（ここでは同義語として使用する）は、予測モデルの開発や検証に用いることができる。研究対象者の組み入れ方法やデータの収集方法などデザインを詳細に記述することは、データの質や、統計解析が適切に実施されたかどうか、予測モデルの一般化可能性に関する詳細な情報を示すこととなる。特異的なバイアスに対する脆弱性はデザインによって異なる。

診断研究は原則的に、診断予測因子（患者背景と検査結果指標）とアウトカム（関心のある対象の状態）の発生や非発生との横断的な関係性にもとづくものである（Box A）。そのような研究にふさわしいデザインは横断研究であり、ある特定の背景をもった患者群が選択され、通常は「関心のある診断したい状態を持ちうる者」（141-151）と定義される。予測因子とアウトカム（参照基準）の測定の間には時間差を生じることがしばしばあり得る。理想的には、この時間差は可能な限り短く、この間にいかなる治療もなされないことが望まれる。測定に短い期間があくことや、似た背景の患者群（これはコホートの定義である）から選択することから、これらの研究を”純粋な”横断研究と呼ぶか、むしろ”診断的な”（横断）コホート研究と呼ぶか”遅延した”{delayed}横断研究と呼ぶかについては議論が分かれる（152-154）。予測因子とアウトカムの測定間隔が長すぎる場合には問題が生じることがある。測定の間に治療が開始された場合は、ある患者においては病状が変化することによって関心のある予測因子とアウトカムについての横断的な関係が崩れてしまう。ここでの主な論点は、ケースと非ケース（コントロール）が、関心のある母集団で起きたケースと非ケースを代表しているかどうか、すなわち、対象としている状態になる可能性を有する個々人かどうかという点である。典型的で進行したケースと健康な非ケースに対して異なる選び方をした研究では、明らかな違反{violation}が起きる（152, 155-157）。そのような対象者の選択は、研究の臨床的な重要性を過大評価することにつながり（158）、予測するためにおこなった測定内容の多くが、しばしば信頼性を失うことになる(157)。

予後研究に相応しいデザインは、縦断的なコホート研究であり、これは前向きと後ろ向きのどちらもあり得る（Box A）(1-3, 58, 103）。特異的{specific}な状態であると診断された、ある特異的な手術を受けた、または妊娠したなどの、特異的な基準にもとづいて、個人はコホートに組み入れられる。この患者組み入れのタイミングをT=0、ベースラインまたは開始点と呼ぶ(9)。次に対象者は、関心のあるアウトカムイベントが発生するか決定されるまで追跡される。

推奨されるデザインは前向きの縦断的なコホート研究である。全ての関連した予測因子とアウトカムが確実に測定されるように管理し（**Figure 2**）それぞれの予測因子とアウトカムを測定するのに最適な方法を使用することによって、欠測値や脱落を最小限に留める。

#### Figure 2. 多変量予測モデル開発研究における予測因子選択



(訳注: TRIPOD公式サイトより修正版をダウンロード 2016/4/4)

多くの研究では、もともと別の目的のためにデザインされ実施されたデータセットを使用し、モデルの開発や検証を行う。そのような研究はもともと前向きの縦断的なコホート研究であるにも関わらず、特異的な予測因子が測定されていないことや、いくつかの予測因子があまり上手く測定されていないことがありうる。Item 13bは潜在的な予測因子の欠測値の数についての詳細について聞いており、item 13aは観察期間中に脱落した患者についてである。

ランダム化試験は、前向きの縦断的なコホート研究のうちの特別なものであり、したがって診断モデルの開発や検証にも使用できる。しかしながら、著者は介入の影響がどのくらいあるかを述べなくてはならない（item 5c）。ランダム化試験のデータを使用したモデルの開発や検証の一般化可能性については、（しばしば）極端な除外基準があることによって、懸念があるだろう（1）。新規の心血管のバイオマーカーが（従来のフラミンガムリスクスコアに追加して）予後にどのくらい寄与するかをみた、ある経験的な解析では、ランダム化試験よりも観察研究から得られたデータの方がより強い影響を認めた（159）。

国際的な協力やデータ共有がより一般的になってきており、多くの研究からの個々の対象者のデータが予測モデルの開発や検証に使用されるようになっている(89, 147,160)。同様に、既存の大きなデータセット（たとえば、国または国際的な調査や登録からの”ビッグデータ”）が予測モデルの開発や検証に使用されるようになっている(139, 161, 162)。これらのデータ源については、対象者はもともと別々のクラスター（別々のコホート、研究、病院、セッティング、地域、国）に属しているためデータはクラスターとして考慮すべき(訳注：対象者が属するサブグループごとにデータの相関がある場合、その点を考慮した解析が必要になる。詳細は引用文献163-164を参照のこと)であり、予測モデルの開発には重み付けが必要である。近年、このような研究対象者のクラスター化に対してメタ解析の手法が提案されるになった (163–166)。このことは、コホートやデータセット、研究、病院、セッティング、地域、国での異なるアウトカムの有病割合（診断）や発生（予後）などを反映したそれぞれの診断関連群{case mix}に対して当てはまる。したがってそれぞれのベースラインの確率やハザード（たとえば、変量切片の使用）にも当てはまる。しかし、このことはまた別の予測因子－アウトカムの関連を反映した診断関連群に対しても当てはまり、無作為予測因子重み付け（回帰係数）を考慮する(163–167)。個々の対象者のデータやほかの”ビッグデータ”の使用は、病院間、国間、セッティング間（**Figure 1**, study type 2b）の予測モデルの開発や直接的な評価（外的検証）の可能性を向上させ、切片や予測因子重み付けの潜在的な違いに対しても適応される (164, 166)。モデルの性能評価において一般的なやり方に、クラスター化されたデータを考慮に入れる拡張も近年提案された(167–171)。

効率やコストの理由から、全コホートの使用よりも患者の抽出が行われる。その例としてケース・コホートやコホート内ケース・コントロール{nested case-control}デザインが挙げられる(172)。関心のあるアウトカムが発生する絶対確率を正しく推定するための解析を行うにはサンプリング方法を考慮する必要があるため、どのように患者（被験者）をサンプリングしたかを正確に報告することが要求される(1, 103, 173–175)。対象者を選択的または除外的に抽出することは、モデルを適応する母集団からのサンプルの代表性が不確かとなり、予測モデルの一般化可能性に影響し得る。

研究デザインやデータソースは、データ収集のセッティングやその本来の目的について関連のある情報を与える。組み入れ基準（item 5b）と組み合わせたセッティングは、読者がモデルを適用したいセッティングに、そのモデルを一般化できるかどうかを判断するのに役立つ。

近年の予測モデル研究のシステマティック・レビューによると、これらの研究では、全ての対象者が連続的に含まれているかなど、サンプルが意図する母集団を代表しているかについて明確に述べられていないことがしばしば見受けられる(34, 59, 84, 93, 176)。

### Item 4b. 組み入れ開始日、終了日、可能であればフォローの終了期間を含む重要な研究の日付を明記する。[D; V]

**例**

この前向きの時間的検証研究は、2007年の3月から6月に英国サットン市の王立マーズデン病院の薬剤開発部門で行われた19のフェーズⅠ試験において連続的に治療された全患者を含み....全ての患者は2008年3月31日まで前向きに観察された (177)。［診断; 検証］

2001年3月から5月に連続して受診した3〜9週間（中央値は5週間）持続する前胸部痛（主訴または付随する症状として）を呈する全ての患者が含まれ....2005年10月から2006年6月の間に前胸部痛を呈する全ての対象患者（35歳以上; n=1249）は、ドイツヘッセ州の74名の参加総合医によって連続的に本研究に採用された。募集期間はそれぞれの診療ごとに12週間行われた (178)。[診断; 開発; 検証]

開発に用いたコホートは2008年5月1日から6月30日にドノスティア病院の内科4病棟に連続的に入院した18歳以上の男女の患者397名であり、除外基準は設けなかった。その後の2009年5月1日から6月30日に同様の条件、つまり同病院の内科4病棟に連続的に入院した18歳以上の男女の患者302名を検証コホートとして採用した(179)。［予後; 開発］

**解説**

研究対象者の募集開始と終了の日付を報告することは、その研究が歴史的状況のなかでどこに位置するかを示す。そのことによって読者はその時期に利用可能な検査や治療、またある予測因子を測定する医療技術力の状態など関連のある情報を推測することが可能となる。これらの日付と対象者の全数との組み合わせによって、選択的な登録がどのようになされたかについても示されるだろう。実際、期間ごと（例えば1年ごと）の登録対象者数の報告は有用である; item 13aを参照。

item 4aで述べたように、ほとんどの診断予測研究において予測因子の測定とアウトカムの測定までの時間は短い。しかし、良い参照基準がないときは、予測因子の測定時から、対象とするアウトカムの状態であるかどうかについて、より信頼できる評価が得られるまで患者は追跡されることがある。このような状況においては、予測因子の測定と対象とするアウトカムの有無についての最終的な評価までに要する期間について、最大または最小でどの程度まで許容されるか、著者は報告しなければならない。予後指標開発研究において、追跡期間はそのモデルの性能を解釈する際に重要となる(item 6aおよび13a)。全対象者において組み入れからフォローの期間は同じとなることがある、その場合研究の期間を具体的に示すべきである。しばしば、研究が終了となる具体的な日付（報告されるべき）まで全ての登録された対象者の追跡は継続される。その場合は、終了日の対象者の状況が確認するよう努め、その日以降に発生したイベントについては無視する。

予測モデル研究のシステマティック・レビューによると、研究の鍵となる日付が常に報告されているわけではない(43, 122, 176, 180)。例えば、乳癌の予測モデルの開発または検証の61の研究のシステマティック・レビューでは、わずか13 (12%)の研究でしか、患者募集の開始と終了および追跡期間の終了日についての全ての情報が報告されていなかった(43)。

## 対象者

### Item 5a. 参加施設の数と場所を含む研究セッティングの重要な要素を明記する（例: プライマリ・ケア、二次ケア、一般人口）。 [D; V]

**例**

我々は、以前のリスク予測アルゴリズム(QRISK1)から修正アルゴリズム(QRISK2)を開発した。我々は、元の解析と類似の手法を用いて、英国の大きなプライマリ・ケア集団を対象とした前向きコホート研究を実施した。我々はQRESEARCHデータベース([www.qresearch.org](http://www.qresearch.org))第19版を使用した。これは、551の開業医によって登録された1100万人の患者の診療録が含まれる認証された大きな電子データベースである(139)。[予後; 開発; 検証］

**Table 2.**も参照。

#### Table 2. 表の例：研究の特徴報告のための表 [診断; 開発; 検証]



**解説**

研究対象者がどこでいつ募集されたかという詳細な記述は、読者がモデルの一般化可能性や有用性を判断し、将来の研究（たとえば、モデルの検証や実行）を実施するためにとくに重要である。” どこでいつ” ということは地理的な場所や暦だけでなく、対象者のデータが収集されたセッティング（一次、二次、三次ケアまたは救急や一般母集団など）や成人または小児の診療かということも意味する。予測モデルをあるセッティングや対象集団からほかのセッティングや対象集団にそのままあてはめることができると単純に考えることはできない。(19, 26, 28, 33)

異なるセッティングには異なる診断関連群が含まれており、これらは通常予測モデルの一般化可能性や正確性に影響する; item 4aを参照(182-187)。“診断群”は予測因子の分布、他の関連する対象者やセッティングの特徴、アウトカムの有病割合（診断）や発生割合（予後）を参照する。そのことはモデルの予測正確性に影響を与えて、異なる予測因子―アウトカム関係を導く可能性につながる。たとえば、よく知られていることとして、二次ケアのセッティングで開発されたモデルをプライマリ・ケアのセッティングに適用すると、通常その予測能力は低くなる(21, 183, 188)。これは単純に、プライマリ・ケア医や家庭医が患者を選択的に二次や三次ケアに紹介しているためであろう。すなわち、二次、三次ケアの母集団においてはプライマリ・ケアに比べて患者背景の幅が狭くなり、進行した病期の患者の割合が大きくなり、しばしばアウトカムの頻度が高くなるということがその理由であろう(102, 189, 190)。

その他のセッティングの違いによる影響として、成人診療における予測モデルを小児診療へ移行できるかという課題がある(102)。

たとえば、成人の全身麻酔下の予定手術における術後の嘔気や嘔吐のリスクを予測する様々な予測モデルが開発されてきた。これを小児において検証を行うと、そのモデルの予測能力は大幅に低下した (191)。

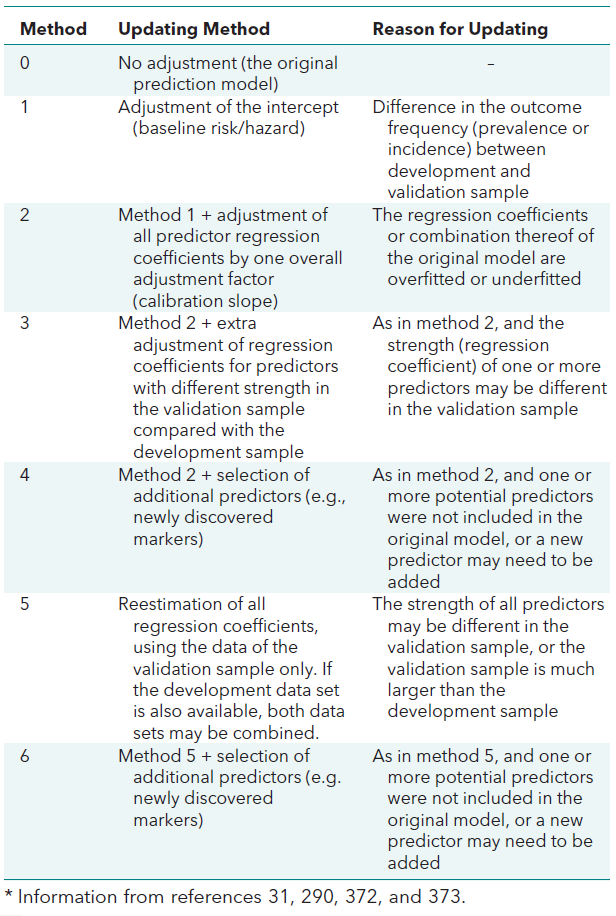
一般に、新たな母集団における診断関連群が、開発母集団の診断関連群の範囲に含まれる場合に、モデルの一般化可能性はより高まる(186)。

しかし、item 10e（Box CおよびTable 3も参照）で述べたように、別のセッティングで開発された既存の予測モデルを、新しいセッティングで当てはまりやすくするために、既存の予測モデルを調整、更新することがある。

読者が診断関連群におけるあらゆる違いやその潜在的な結果（item 5b）に対する洞察を深めることができるように、開発や検証のサンプルの対する研究の背景の鍵となる要素をまとめた表を示すことを我々は推奨する(192)。加えて、検証研究においては、開発用サンプルと検証用サンプルのそれぞれのまとめの表を示すことを我々は提案する。

心不全の予測モデルの開発または検証をおこなった48の研究からなるシステマティック・レビューにおいて、10(21%)の研究が施設数の情報を示していなかった(180)。

#### Table 3. 既存の予測モデルを更新する異なる方法の概観\*



### Item 5b. 対象者の組み入れ基準を記述する。[ D; V]

**例**

1987年から2002年にUMC ベンジャミン フランクリン ベルリンとUMCマンハイムの皮膚科において、192名の皮膚リンパ腫の患者が評価された。欧州がん研究・治療機構の皮膚リンパ腫分類（菌状息肉腫、セザリー症候群、そして稀な変異を含む）によって定義される皮膚T細胞リンパ腫(CTCL)と診断された患者は86名であった。...CTCLの稀な変異、類乾癬、皮膚偽リンパ腫および皮膚B細胞リンパ腫の患者は研究から除外した...

病期分類は菌状息肉腫共同班{mycosis fungoides Cooperative Group}のTNM分類法に従っておこなった。セザリー症候群は、紅皮症を伴い、末梢血のセザリー細胞が1000/mm3（以上という国際皮膚リンパ腫学会(ISCL)の基準に従って診断した(193)。[予後; 開発］

組み入れ基準は12歳以上および受傷後7日以内である。12歳未満の患児は我々の病院から半マイルの距離にある近隣の小児病院で診療され、我々が救急で受け入れるのは主に12歳以上患児である。そのため、12歳をカットオフとした。これは、18歳以下を除外としたStiellの原著と異なる点である。除外基準は、妊婦、コンサルテーション時点の精神状態変化、X線検査を既に受けている紹介患者、再来患者、多発外傷、熱傷や擦過傷、裂傷、刺創のような単独の皮膚損傷(194)である。［診断; 検証］

**解説**

組み入れ基準を記述することは予測モデルの潜在的な適用性、つまり一般化可能性を理解するうえで重要である。誰が研究対象者となり、誰がならなかったかを説明する選択過程は、研究の結果がもたらされる対象者や予測の一般化可能性に関連している。

検証研究にとって、組み入れ基準がモデルの元の作成研究で使用されている基準と違うのか同じなのかを報告することは有用である。上記の例(194)では、元々カナダで作成された足関節の骨折を識別する多変数の診断予測モデルを、アジアで検証している。著者らは選択の詳細について説明し、元の開発研究で使用された方法と対比をさせている。

もし、欠測データが理由で組み入れられるべき対象者が除外される場合は、そのことを報告で明記しなければならない。単に欠測値（予測因子およびアウトカム）があることを理由に対象者を除外し、予測因子とアウトカムのデータが完全に観察できたもののみを解析の対象に制限することで、重大なバイアスが起こる可能性がある(195-201)。データに完全にランダムな欠測{missing completely at random}が起きるのではなく、むしろ選択的に欠測が起こるためにバイアスが生じるのである（item 9）。

### Item 5c関連があれば治療の詳細を記述する。[D; V]

**例**

多施設国際共同臨床試験（Action in Diabetes and Vascular disease: preterax and diamicron MR controlled evaluation)(ADVANCE) では糖尿病患者の心血管系リスクの新しい予測式を開発{derivation}する. . . . ADVANCEはアジア、オーストラリア、ヨーロッパ、カナダといった20カ国215施設で行われた2型糖尿病のハイリスク患者11140人を対象とした微小血管あるは大血管系のイベントの発生に対する血圧（perindopril indapamide　対　プラセボ）と血糖コントロール（gliclazide MRに基づいた集中的な介入　対　標準ケア）の効果を検証する2×2{factorial}ランダム化比較試験であった. . . .考慮された予測因子は糖尿病診断年齢、糖尿病診断後の期間、性別、収縮期血圧、拡張期血圧、平均心房圧、脈圧、総コレステロール、HDL、非HDL、トリグリセリド、body mass index、ウエスト径、ウエスト・ヒップ比、血圧降下薬（つまり高血圧の治療）、スタチン使用、現在の喫煙、糖尿病性網膜症、心房細動(既往あるいは現症)、対数変換したアルブミン／クレアチニン比(ACR)と血清クレアチニン(Scr)、ヘモグロビンA1c（HbA1c）、空腹時血糖値、ランダム化された治療（降圧薬および血糖降下薬内容）であった(145)[予後; 開発; 検証]。

**解説**

予後の研究におけるコホートはいくつかの健康状態によって定義づけられる(202)。多くの予後に関する研究において、対象者は予防的あるいは治療的介入をフォローアップ期間の前あるいは開始時に受ける。このことは、予後に影響をあたえるだろう。効果的な治療を受けることは、アウトカムの発生する可能性を下げることで、予後を改善するだろう (203)。

特定の健康状態で治療を受けたことがない対象者に限定して将来のアウトカムについて『純粋なベースライン』の予後モデルを作るということはほぼ不可能である。しばしば対象者は何らかの治療を受けている。予後モデルがランダム化試験のデータに基づいている場合全ての研究対象者は手術のような治療については同じものを受けているか、治療はランダムに選択されるのが理想的である; item 4a参照(1,204)。特定の介入を受けた患者に対してのみ、開発され、検証された予後モデルもある(205)、しかし、それでも、複数の介入を受けることで結果が変化するかもしれない。

ランダム化試験のデータが使用されたとき、効果的な介入がある場合は特に、介入群とコントロール群で別々の予後モデルが適合するかもしれない。代わりに、介入群とコントロール群の両方から開発したモデルでは、治療は予測因子として含むことになる(item 7a)。治療と他の予測因子の交互作用(item 10b)が、異なる治療戦略のもとで、異なる予測を可能にするために研究されうる(1,4)。ここで大事なのは介入の予防的あるいは治療的効果ではなく、アウトカム予測に独立してどの程度寄与するかである。しかしながら、多くの場合、介入の予測効果は、例えば年齢、性別、病期などの重要な予測因子と比較してかなり小さく、治療はモデルから除外されるかあるいは、予測因子を選択するプロセスの過程で除かれる。

非ランダム化研究における重大な懸念として、受ける治療にばらつきがあるということだけでなく、個々人への治療選択が統計モデルに含まれている予測因子と同じ要因によって影響を受けている可能性が挙げられる。(206)。ランダム化試験のデータと同様、治療はモデルの予測因子として考慮されるが、治療が他の予測因子から影響を受けている時の予測モデルに与える効果は、簡単に判断することはできない。

前述した意見はフォローアップ期間開始時より前に受けた治療と関連している。フォローアップ期間開始後に受けた治療を考慮に入れるには、予測モデルの研究において使われることがほとんどない、非常に凝ったモデルを必要とする(207)。

現在の治療が他の潜在的な予測因子の代わりとして用いられる場合はかなり異なった状況となる。上述の例では、降圧薬あるいはコレステロール降下薬の使用を高血圧あるいは脂質異常症の代替変数として、それぞれ心血管リスクモデルに組み入れている(17,208)。予測モデルにおいてこのようなアプローチがどのような結果をもたらすかは未だ十分に分かっていない。上述のような観点から、どの介入を研究対象者が受けることでアウトカムの発生率に影響があるかを知ることは、予測モデルを開発、検証することの両方にとって重要である(203)( item 13b)。

治療についての問題は、ほとんどの診断予測モデル研究においてあまり問題にならない。（使用される）横断研究のデザイン では予測因子とアウトカムが同時に記録されるためである。(Box A)。しかしながら、予測因子とアウトカムの評価の間に時間差(たとえば、アウトカムの測定にフォローアップ情報基づいているとき[209])のある場合がある。その結果、予測のタイミングとアウトカム測定の間に受けたどのような治療も予測に関連のある情報となるため、報告されるべきである。

21の心血管リスクスコアに関する最近のレビューでは、アウトカムに影響のある介入が不明であり、先行する治療の報告は不完全であったということがわかった(203)。

### Item 6a　どうやって、いつ評価されたかを含めて、予測モデルで予測されるアウトカムを明示する。[D; V]

**例**

関心のあるアウトカムはあらゆる死亡、冠血管性心疾患関連死亡、そして冠血管性心疾患イベントであった。それらのアウトカムを測定するために、研究対象者は、年に一度の電話インタビュー、3年毎の研究関連機関での診察、 ARIC関連病院の調査、死亡診断書の参照、医師の質問紙票、検察医の検察レポート、そして情報提供者との面談といった様々な方法で長い時間をかけてフォローされた。フォローアップは登録時(1987年から1989年)に始まり、2000年の12月31日まで続いた。冠血管性心疾患関連死亡には冠血管性心疾患に関連した入院および非入院死亡を含んだ。冠血管性心疾患イベントは入院の必要であった確診または疑いの心筋梗塞、致死性冠血管性心疾患、心臓手術(冠動脈バイパスグラフト、冠血管形成術)、あるいは3年毎のコホート検査で心電図の変化が存在すること、と定義された。イベントの分類は他の場所に詳細を記載している[参照](210)。[予後; 開発]

尿路感染症の診断確定は、排尿サンプル内で1種類の微生物が108コロニー形成単位(colony forming units: cfu)/L以上、カテーテルサンプル内で1種類の微生物が107 cfu/L以上、あるいは恥骨上膀胱穿刺サンプルにおけるあらゆる微生物の増加と定義された。可能性の高い尿路感染症は排尿サンプル内で1種類の微生物が107cfu/L以上、カテーテルサンプル内で1種類の微生物が106cfu/L以上、排尿サンプル内で2種類の微生物が108cfu/L以上、あるいは、カテーテルサンプル内で2種類の微生物が107cfu/L以上と定義された。[診断; 開発; 検証]

患者の診療記録と医師記録が臨床アウトカムを決定するために再検討された。患者は多くの場合、術後1年以内の最低でも3−4ヶ月毎に、術後2−3年後は年2回、それ以降は年1回診察を受けていた。すべての患者でフォローアップ検査としてCTによる放射線画像検査を実施していた。身体診察に加え、検体検査、経静脈尿路造影、膀胱鏡検査、尿細胞診、膀胱洗浄、そして骨シンチが必要とされた場合に実行された。

局所再発は手術腔における再発と定義され、遠位臓器での再発を遠隔再発とした。臨床アウトカムは膀胱切除術日からCTで最初に再発と記載された日、死亡日、患者が再発を経験せずフォローが終了した最終日までの日数で定義した。[予後; 開発]

乳癌確認: 乳癌の偶発的な診断は1997年から2005年にかけての3年毎のフォローアップにおける自己申告で確定した。我々は家族、アメリカ合衆国郵便公社、そしてNational Death Indexから死亡を測定した。我々は1084例の偶発性乳癌を確認し、1007例(93%)は医療記録あるいはベースライン時において96％の対象者が住んでいた24の州におけるがん登録で確認された。[予後; 検証]

**解説**

診断モデルでは、アウトカムはT0において特定の診断したい状態が存在するかしないか、となる**(Box A)**。この診断アウトカムは、ある状態があるかないかを確定するための方法として最も利用され広く受け入れられている、いわゆる参照基準を用いることによって決定される(214)。参照基準の選択根拠は示されるべきである。参照標準は多くの形態を取り得、ひとつの検査の場合もあるし、検査の組み合わせ、あるいは専門/アウトカム委員会を用いた、意見の一致を基本とするアプローチを含んだ他の方法の場合もある。参照基準検査は検体検査、放射線、関節鏡、血管造影、あるいは病理学的評価などがありうる(訳注: 原文はbeとなっているが、may beの間違い)。

もし関連があるのであれば、血液あるいは尿サンプリングの方法、測定法と画像検査法や技術、そしてその定義を、診断したい状態の存在(あるいは重症度)を定義するのに用いられたあらゆるカットオフを含んで明示する必要がある。同様に診断を確定するために、検査結果を結合(結合参照基準)した方法も明示すべきである(215-217)。標準的な定義と閾値を用いない理由は明示するべきである。もし複数のアウトカム評価者(合意を基本とするアウトカム評価委員会を用いるときのように)が用いられた場合、最終診断を確立する方法(例えば多数決)は記述されるべきである(215,216)。

診断モデルに関する研究において、予測因子を評価してから後に基礎的状況が変化することによって、バイアスを生じる可能性があるため、アウトカムの評価と関連のある予測因子の評価のタイミングは記述されるべきである(**Box A**)。さらに、予測因子とアウトカムが評価される順番については明確に記載されるべきである(マスク化されていない評価に関連するバイアスの可能性についてitem 6と7も参照)。

理想的には、診断アウトカムはすべての対象者が同じ参照基準を用いて検証されるべきだが、それは常に可能なわけではない。例えば、患者に、もし１回あるいはそれ以上の診断指標にて陽性となっていない場合、侵襲的参照基準を無理に行なうのは非倫理的とみなされるかもしれない。ふたつの状況が次に起こりうる。ひとつは不完全な検証{partial verification}で、これはアウトカムデータが、参照基準の結果がない対象者の中で、アウトカムが完全に欠測(item 9)している状態である。もう一つは、異なる検証{differential verification}で、これは推奨される参照基準のデータが欠測している患者が、他の異なる、一般的に低い精度の参照標準を用いて評価されたときに生じる(218,219)。

例えば、癌検診の研究において、病理(参照基準)結果はいくつかの検査で陽性となった対象者のみで可能となりやすい。残った対象者については、代替となる参照基準が、(後の検証で)明らかとなる検査の測定時、癌を発見するのに充分長期とみなされたフォローアップ期間となってしまう可能性がある。しかしそれは新たな癌の発生に十分な期間ではないかもしれない。アウトカムを不完全にあるいは偏りをもって検証してしまう場合の基準と方法については、いわゆる不完全/鑑別検証バイアス{verification bias}の可能性を評価できるよう充分に記載されるべきであり(156, 218–220)、それらの検証バイアスを調整するための方法は考慮されるべきかもしれない(219)。

予後モデルにおいて、一般的なアウトカムは死亡(全死亡あるいは特定の原因による死亡)、致死的でない合併症あるいはイベント(たとえば心筋梗塞、癌の再発、病気の進行、あるいは病気の発症)、患者中心のアウトカム(たとえば症状、機能、QOL)を含む(2)。アウトカムの組み合わせも同様に用いられ、例えば、がんの研究における無病生存期間のとして関心のあるイベントは局所再発、同一臓器内再発、遠隔転移、そして死亡(どれが最初に起ころうと)を含んでいるだろう(221)。

すべてのアウトカムは曖昧さをのこさず定義されるべきである。もし標準的な基準(例えばICD{ International Classification of Diseases }コード)を用いた場合、そのことは全ての変更点とともに記述され、引用をされるべきである。研究プロトコルあるいは先行研究で提供されていた技術的な詳細はわかりやすく参照されるべきであり、理想的には入手できるようにするべきである。

予後の研究は対象者を一定期間にわたりフォローし、アウトカムが開始時点(T0) (たとえば、診断日や手術日; **Box A**参照)よりも後でいつ生じたかを記録する。研究によってはすべての対象者に対して一定期間内アウトカム(例えば全生存)の評価をするものもあれば、フォロー中の予め規定された時間(たとえば５あるいは１０年間の心血管疾患リスク)を評価する場合もある。それらの期間についても明確に報告されるべきである(222)。

同様に、フォロー期間中のアウトカム評価の頻度についても明確に報告されるべきである。アウトカム発生や脱落を見つけるために使用されたデータ・ソースは明示されるべきである。例としては死亡レジストリデータ、病院記録、癌レジストリデータ、診療評価、画像検査、検体検査といったものがある。特定の原因による死亡のようなアウトカムに関して、死亡原因が決まった過程は明確に説明されるべきである(たとえば判定委員会やエンドポイント委員会)。そのような委員会の構成とメンバーの専門性は簡潔に記載されるべきである(216)。

癌予後モデルの発達について報告している47の最近のレビューにおいて、アウトカムの定義は研究の40％において不十分であると結論されている(54)。30%において、『死亡』が癌死亡かすべての原因による死亡かはっきりしなかった。どのイベントが無病生存期間の定義に含まれたかという点についても同様につじつまが合わなかった。

### Item 6b予測されるアウトカムの評価に際して、マスク化を行うために実施した行動について報告する。[D; V]

**例**

重篤な細菌感染症の可能性が高い症例は、2人の小児科専門医(小児感染症と呼吸器内科の経験がある)と、肺炎の場合には放射線科医によって構成された最終診断委員会によってレビューされた。細菌感染症の有無[アウトカム]は臨床情報[研究における予測因子]をマスク化した状態で決められ、意見の総意に基づいたものであった(211)。[診断; 開発; 検証]

肝生検は18Gあるいはそれ以上の太さの針で穿刺を行なった。5本以上の門脈が含まれている検体をヘマトキシリン−エオジン染色とトリクローム染色にて染色した。生検はMETAVIRグループによって作成されたスコアに従い、患者の臨床的特徴や血液検査結果がマスク化された状態で2人の専門肝臓病理医によって評価された。30の生検が両方の病理医によって得点化され、観察者間の一致率はカッパ統計量を用いて計算された(223)。[診断; 開発; 検証]

プライマリアウトカム[急性心筋梗塞の冠動脈血管再開通術、心臓死、あるいは30日以内の原因不明の死亡]は予測因子の変数をマスク化された研究者によって確定された。診断が確定されなかった場合は・・・心臓専門医がすべての臨床データをレビューし、アウトカムの診断を決定した。すべての陽性と10％のランダムで選択された陰性アウトカムは、標準化されたデータフォームについてマスク化された二人目の共同調査者によって確認された。相違点は意見の一致を持って解決された(224)。[予後; 開発]

**解説**

予測モデル研究において、アウトカムは予測因子がマスク化された状態で評価されるのが理想である。予測因子はアウトカム評価に影響をあたえるかもしれず、予測因子とアウトカムの間の関連を推定する際にバイアスを生じる可能性がある(148,209,225,226)。このリスクは、例えばすべての原因による死亡、あるいは帝王切開といった客観的なアウトカムでは明らかに小さい。しかし、疾患特異的な死亡のような解釈を必要とするアウトカムの評価をする際には関連が高くなる。

確立された参照基準による評価がもともと難しく、参照基準を欠くようなアウトカムがある。研究者はすべての入手可能な各々の患者の情報(予測因子からの情報を含む)を、アウトカムの有無を真に決定するために、使用したいと考えるかもしれない。診断の研究において、このアプローチはコンセンサス診断として知られ、予後研究における裁定委員会あるいは介入研究におけるエンドポイント委員会が代表例である(item 6a)(149)。もし研究目的が特定の予測因子を既存のモデルに付け加える価値を評価することや、競合モデル(たとえば、複数のモデルの検証)の性能を比較するためということであるなら、予測因子の増分価値を過大評価するのを防ぐ、あるいは1つのモデルがもう一つのモデルよりも偏って好まれることを防ぐため、マスク化されたアウトカムの評価の重要性は増してくる。

そのため、研究者は注意深く考慮する必要がある。どの情報がアウトカムの評価者に利用可能であるかを明確に報告し、適切な場合は、アウトカム評価をマスク化するためのあらゆる具体的な取り組みを記録すべきである。しかしながら、アウトカムの評価がマスク化されているかどうかの評価に際して、情報が不足していたと報告するシステマティック・レビューが多い(34,227)。

## 予測因子

### Item 7a. どうやって、いつ評価されたかを含めて、多変量予測モデルの開発または検証に使われたすべての予測因子をはっきりと定義する。[D; V]

**例**

各患者から下記のデータを抽出した: 性別、AST { aspartate aminotransferase } (IU/L)、ALT {alanine aminotransferase }(IU/L)、AST/ALT比、総ビリルビン(mg/dl)、アルブミン(g/dl)、トランスフェリン飽和率(%)、MCV {mean corpuscular volume }(µm3)、血小板数(×103/mm3)、プロトロンビン時間(秒). . . .全ての検査は肝生検前の90日以内に行った。反復して検査している場合は最も生検日に近い日の結果を用いた。生検後の検査結果は用いなかった(228)。[診断; 開発]

AMI[急性心筋梗塞]の死亡率予測モデルを開発するにあたり、年齢、性別に加え43個の潜在的な変数の候補が検討された。変数の候補は以前に行われたCalifornia Hospital Outcomes Projectと、Pennsylvania Health Care Cost Containment Council AMI “report card” projectsで使用されたリスク要因の一覧から選択された。各合併症についてはOMIDにおける15の二次診断分野から適切なICD9コードを用いて作成した。Ontarioの退院データは、米国で使用されているICD-9 CMコードではなくICD-9コードに準じているため、米国コードが省略されたものである。この2つのプロジェクトで使用されているリスク因子にはICD 9コードに含まれないものがあり（例、梗塞のサブタイプ、人種）、そのため、それらは本研究の分析に含まれなかった。43の併存症の頻度が計算され、有病割合が1%未満の合併症は除外された。また、リスク因子のうち臨床的に急性心筋梗塞の死亡率の予後因子としては考えにくいと著者らが感じたものについては除外した (185)。 [予後; 開発; 検証]

それぞれのスクリーニングは約3週毎の外来受診を2回含んでいる。試験対象者は人口統計属性、心血管疾患の既往、腎疾患の既往、喫煙歴、内服薬（糖尿病薬、降圧薬、脂質降下薬）について質問紙に回答した。降圧薬の種類などの情報については地域の薬局からデータを収集し把握した。1回目と2回目の来院の際にそれぞれ10分間と8分間、右腕で毎分血圧測定を行った。血圧測定器は Dinamap XL Model 9300 series (Johnson & Johnson Medical Inc., Tampa, FL)を用いた。収縮期血圧と拡張期血圧については各来院における測定の最後の2回の平均を用いた。身体測定を行い、採血は空腹時に行った。総コレステロール値、血漿血糖値については標準的な方法で測定した。血清クレアチニン値については同時再現性係数が0.9%、測定間再現性係数が2.9%であるドライケミストリー(Eastman Kodak, Rochester, NY)を用いて測定した。eGFR [estimated glomerular filtration rate]は性別、年齢、人種、血清クレアチニン値から推定するModification of Diet in Renal Disease (MDRD) 研究の式を用いた。さらに、対象者は2日間連続で24時間蓄尿を行った。尿中アルブミン濃度はネフロメトリー(Dade Behring Diagnostic, Marburg, Germany)で測定し、尿中アルブミン排泄量 (UAE, urinary albumin excretion)は2回の24時間蓄尿の平均値を用いて算出した。食塩とタンパク質の摂取量の代用として24時間蓄尿におけるナトリウム排泄量と尿素排泄量をそれぞれ用いた (229) 。[予後; 開発]

**解説**

予測因子は典型的には対象者の人口統計属性や医療記録、身体診察、疾患特性、検査結果、以前の治療情報を用いる (1)。予後因子については連続変数であればその測定単位を記述し、カテゴリー変数であれば全てのカテゴリーを記述し（カテゴリーを統合したかどうかを含めて）、厳密に定義すべきである。定義することにより読者や他の研究者による試験の再現性を保証し、さらに重要な点としては予測モデルの妥当性の検証及び実用性を保証する。もし可能であれば、適切なサンプリング、検査室そして画像検査法についても明示すべきである。その中には、明示された予後因子の有無(または重症度)を決めるカットオフ値や、予後因子を統合する際のやり方（例えば平均血圧など）といった内容を含める。

そして、著者らはいつどのように予後因子が測定されたのかも説明すべきである。全ての予後因子は試験開始前もしくは開始時に測定されるべきであるし、予後予測モデルを使用する時点で知ることができる因子であるべきである。(1,230,231)血液や組織のサンプルは試験開始前もしくは開始時に採取されるが、分析されるのは後になるかもしれない; その場合に重要な問題は、いつサンプルが採取されたかということと、予後因子がいつ使われるのかということである。試験開始後に測定された因子はアウトカムをより反映するかもしれないが、時間依存性解析を用いない限りは予後因子とならない(232)。しかしながら、予測モデルにおいて、フォローアップ期間に測定された予測因子を扱うための統計学的手法(233, 234)が用いられることはめったにない。予測因子の測定方法（アッセイの方法や検査室での測定方法を含む）の報告は完璧かつ透明性を持った様式で行い、測定を再現できる程度にまで、またそのような予測因子を含んだ予測モデルの一般化可能性を評価できる程度にまで詳細に行う。

予測モデルを開発している多くの研究では、多くの予測因子が集められ、統計解析可能となっている(**Figure 2**）。しかしながら、解析可能な予測因子がより増えることで、関連が弱く情報価値のない因子を誤って最終モデルの中に選択してしまう危険が増大し、いわゆるモデルの過剰適合や過度の期待を引き起こす（特に小さいデータセットの場合）; item 8参照。

さらに、少ない予後因子からなる予測モデルの方が、多くの予後因子からなる予測モデルと比べて、臨床現場では適応しやすい。そのため分析前もしくは分析途中でモデルに入れる因子を減らす作業はしばしば必要となる(2,235); item 10b参照。

モデル作成に使われた予測因子のセットから、どの予測因子を省く場合も、その理由を明確に報告すべきである(**Figure 2**)。

全ての予測因子、解析された予測因子の数、そしていつどうやってそれらが選択されたかについて、しばしば不十分にしか報告されていないことが、最近のシステマティック・レビューで明らかになった (34,43,45,53,54,73,80,81,87,182)。29の生殖医療の予後予測モデル研究のレビューによると、34％の研究で予後因子の記述を適切に行えていなかった(80)。

### Item 7b. 予測因子の評価に際して、マスク化を行うために実施した行為について報告する。 [D: V]

**例**

臨床データと心エコー評価を知らされていない一人の評価者が定量的なMRIの分析を行った。[心不全の有無を診断するに際して、臨床データを超えたMRIの増分価値を具体的に定量化することを目的とした](236) [診断; 開発; 増分価値]

[他の]予測因子や患者アウトカム[胸痛発症から30日以内の非致死的あるいは致死的心血管疾患の合計と総死亡]についてマスク化した後、2つの救急専門医が全ての心電図[研究における予測因子の１つ]を構造化された標準フォーマットを用いて分類した. . . (224)。[予後; 開発]

予測因子と患者アウトカムを知らされていない研究者が、最新の標準化された報告ガイドラインに従って、全ての心電図を構造化されたフォーマットを用いて評価及び分類した。標準化されたデータ収集フォームについて知らされていない２人の研究者が患者アウトカムを確認した。その研究者たちは検査値、放射線画像、心負荷試験、心臓カテーテル所見、さらに電話で得られた30日間のフォローアップ情報を与えられた。[診断; 検証]

**解説**

もし評価者がある予測因子の評価をする際に、アウトカムや他の予後因子の情報についてマスク化されていなかったら、その評価に何らかの影響をもたらす可能性がある (1,225,238,240)。アウトカムの評価者をマスク化することと同様に(item 6b)、特に画像診断や電気生理学検査、病理診断などの主観的な評価や判断が必要な予測因子について、評価者をマスク化することが重要となる。一方で年齢、性別、量的に評価できる検査値などの観測者の主観的解釈とは独立したものであればその必要性は大きくない。

***アウトカム情報のマスク化***

予測因子の評価は常にアウトカムの情報なしに行うべきである。アウトカムの情報が予測因子の評価に間接的に含まれる、もしくは混入すると、アウトカムと予後因子の関連を強めることになる (1, 225, 239)。

予後の研究で一般的に行われているように、アウトカムが予測因子の後に測定されるようなフォローアップ研究では、もともと予後因子の評価者にアウトカム情報は分からない。したがって、アウトカム情報が予測因子評価に混入した際に生じるバイアスは、とりわけケース・コントロール研究及びアウトカムと予測因子の測定時期が近い横断デザインの研究で生じることになる(225)。それ故にこのバイアスは診断モデル研究においてより生じやすい。したがってアウトカム情報が予測因子(もしくは診断指標)の結果を解釈する際に手に入るかどうかは明確に記載しておくべきである。

***他の予測因子の情報のマスク化***

予測因子の解釈が必要となる評価者は、他の情報(例えば既往歴や身体所見を評価する際に得られた情報)も提供されるかもしれない。予測因子の評価をする際のアウトカム情報のマスク化と違い、他の予測因子に対するマスク化それ自体が良いとか悪いというものではない。その適切さに関しては、リサーチクエスチョンと特定の予後因子の潜在的な臨床応用に依存する。(209, 225, 226)もし日々の臨床である予測因子の情報を踏まえて別の予測因子を解釈することが多い場合は、ある予測因子を知った上で、次の予測因子を解釈するように明確に計画することもある。例えば追加の画像検査や電気生理学検査から得られた予測因子は、一般的に病歴や身体所見の情報とともに解釈される。

また、いずれにしても、もし研究の目的が臨床の場において何らかの形で知られている特定の予測因子に対する増分価値を定量化することであれば、前の予測因子の情報を後の評価者から隠すことは不当となるかもしれない。しかし、もし研究の目的がある予測因子もしくは検査が他の予測因子もしくは検査を置き換えることができるか、ということを定量化することであれば(例えば、肺の癌病変を検出する従来のCTスキャン法に、PET-CT検査が取って代わるかどうか)、両観測者に対し互いの結果をマスク化することが、両者の解釈が混ざることを予防するために必要である。(225,239)

非マスク化は両者の読み、そしてその結果を似通ったものにする傾向がある。

そのため、研究の目的と、臨床のどこでどのように予測因子が使われるのかという点に関して、どの予測因子の評価が他の予測因子情報にマスク化されているかについて、どんなことでも報告すべきであろう。

予後因子のマスク化については、行われていないかもしくは報告されていないかのどちらかであることが、多くのシステマティック・レビューで示されてきた。(3,58,67,95,241)例えば、小児科領域の予後予測モデル開発を行った137の研究のうち予後因子のマスク化についての検討が報告しているものは47%しかない。

## サンプルサイズ

### Item 8. サンプルサイズがどのように決まったか説明する。[D; V]

**例**

得られた決定ルールの感度の推定精度からサンプルサイズを推定した。以前の決定ルール研究の通り、100%の感度で95%信頼区間の下限が97.0%となる条件を満たすよう120のアウトカムイベントを事前に規定した。そして臨床現場の救急医に最も有益になるように、救急外来の外(院内もしくは救急外来から帰宅した後)で生じた少なくとも120のアウトカムイベントを含めることを目指した。「オタワ病院」の診療の質データでは、胸痛を主訴に救急外来を受診した患者の10%が30日以内にアウトカム基準を満たした。これらのイベントのうち半分は入院後もしくは救急病棟帰宅後に生じると推定した。事前に規定されたサンプルサイズは2400人と推定された。(224) [診断; 開発]

我々のサンプルサイズ計算は主要な目的(すなわち手術前の冠動脈造影CTが、他の臨床的な因子以上に予後予測としての価値があるかどうかを決定すること)に基づいている。2つの目的のうち、この目的に対しては大きな対象者人数が予測モデルの安定性を確保するために必要となった。VISIONパイロット研究と以前の非侵襲的な心機能検査の研究に基づき、本研究における重要な周術期心疾患イベントの発生率は6%と想定した。表 2は異なったイベント発生率と変数ごとに必要なイベント数から、４つの変数を多変量解析する上で必要なサンプルサイズの一覧を表している。表が示している通り、もしイベント発生率が6%であれば安定した推定を行うために1000人の対象者が必要である。もしイベント発生率が4%であった場合、最大1500人の対象者が必要となるかもしれない。サンプルサイズは1500人を目標としたが、これは1000人あたりのイベント発生率によって変わる可能性がある。(242) [予後; 開発]

データベースにおける全ての利用可能データは検出力と結果の一般化可能性を最大にするように使用された。(243) [診断; 開発]

全てのコホート研究が現在進行形であるため、型通りのサンプルサイズ計算は行わなかった。また、リスク予測モデルにおいて作成研究と検証研究に必要なサンプルサイズを推定するための、一般的に受け入れられている方法はない。モデル開発研究では各変数の候補毎に最低10イベントが必要となり、モデルの検証では最低100イベントが必要としている文献もある。予測モデルの開発及び検証を行っている多くの研究は小規模であるため、本研究のような大規模共同研究を行い、回帰モデルによる安定した推定を行うことで、他の集団にも一般化しやすい結果を得ることが一つの解決策になるかもしれない。本研究のサンプル数とイベント数はサンプルサイズを推定するどの方法で期待される数よりもはるかに多い、それゆえ非常に頑健な推定をもたらすことが期待される。(147) [予後; 検証]

検証のために必要なサンプルサイズは期待アウトカム(全ての腹腔内損傷) が100人必要と設定した。この計算方法は以前行われた臨床予後予測モデルの外的検証で記述された統計学的推定に従っている。以前の我々の研究と同じく、本研究のサンプルにおける腹腔内損傷の発生率は10%と推定し、そのため1000人の対象が必要であると計算された。(244) [診断; 検証]

**解説**

予測モデルを作成するにあたっては、十分なサンプルサイズがあることが重要であるというコンセンサスはあるが、何人いれば「十分か」を決める方法については、明らかではない。すべての医学研究について、より大きなサンプルサイズはより精度の高い結果を生む。バイアスがなければ、大きなサンプルサイズはより信頼のおける結果を生む。大事なことは、予測研究(開発と検証)においてアウトカムの数が有効なサンプルサイズを決めるということである。二値変数やイベント発生までの時間がアウトカムとなる場合、有効なサンプルサイズは、2つのアウトカムの頻度より小さい。全体の人数という意味で大きなサンプルサイズは、もしほどんど実際のアウトカムが起こっていない場合は不十分ということになるだろう。

しかしながら、しばしばデータセットはあらかじめ潜在的な予測因子とアウトカムが測定された完全で大きなコホートから入手できることがある。そのため、明確なサンプルサイズの設計をしたかどうかに関わらず、コホート全体のデータを用いる意味はあるだろう。もしそうであるなら、そのような状況であることを恣意的な事後のサンプルサイズ設計で正当化せずに、明らかにするべきである。

***開発研究***

Item 10b以下で議論したように、同じデータセットで作成と正確性の検証がなされた場合、モデルの性能は過剰に見積もられる(23)。この問題はサンプルサイズが小さい場合に最も大きくなる (25, 32, 112)。モデルにおける過度の期待の度合いについては、内的検証およびシュリンケージの手法を用いることで、推定され調整可能ではあるが(item 10bにて議論)、まずは大きなサンプルサイズを確保することの方が良い。これらの懸念は、予測因子の選択が実施されなかった場合にも当てはまる。しかしながら、モデルの予測因子が数多くの予測因子候補の中から選び出されている場合（Figure 2）、中でも強力な予測因子が無い場合には、特に大きな問題となる。少ないサンプルサイズからは、不適切な予測因子を選ぶリスクが高くなる（過剰適合; item 10b）そして、重要な予測因子を選び損ねるリスクが高まる（過小適合）(25, 26, 32, 112)。

いくつかの経験的な調査にもとづき(245, 246)サンプルサイズの経験則{a rule of thumb}が広く採用されてきた。そのルールとは、一つの変数、より正確には多変量解析で推定される変数ごとに、最低10の変数毎のイベント(EPV, events per variable)があることである。しかしながら、10はあまりに厳格すぎる(247)、もしくはゆるすぎる (25, 32, 248, 249)ことが示唆されている。加えて、EPVはこういった判断にベストの基準ではないだろうとされている(250)。原則として、サンプルサイズは与えられた精度のもとで推定されるモデルの性能が明確な基準に合致できるようにするために選ばれるものである。モデルの性能の評価で用いられる方法には、c-index、R2、Brierスコア、感度と特異度、その他といったものがある（251-253）。

実際には、研究者はしばしば目の前にあって使用可能なデータセットを用いるしかない状況にある。EPVのような評価法はしばしば記述的なものに過ぎないが、解析される予測因子の数はEPVをコントロールするために削られるであろう(**Box C**)。上に述べたようなアプローチ法にもとづき、サンプルサイズを統計学的に事前に決められるのは、前向きの予測モデル作成研究のみである。

論文著者はサンプルサイズがどのように決められたのか説明すべきである。統計的な考慮に基いているのであれば、詳細を述べるべきである。しばしば、サンプルサイズは、時間、存在するデータの利用可能性、コストといった実務的な判断から決められる。こういった場合、予測因子の数、もしくは予測モデルに対する評価法との兼ね合いでサンプルサイズが十分かどうかを議論することが助けになる。

***検証研究***

検証研究には具体的なゴールがある: 　既存のモデルの性能を他のデータで定量化することである（Box CとFigure 1）。検証研究に必要なサンプルサイズについてはよく分かっていない。研究者をガイドする経験的証拠は足りていない。それゆえサンプルサイズはしばしば、利用可能なデータによって決まる、しかし一部の場合において統計的な基礎によりサンプルサイズを選ぶことができる。

検証研究のためのサンプルサイズを選ぶのに研究者を助ける限定的な証拠としては、最低限100のイベントと100の非イベント数が必要とされている（112, 254）、一方で250以上のイベントがあることが好ましいと提案されたこともある(2)。しかしながら、こういった提案は統計的な仮説検定のフレームワークを採用した限られたシミュレーション研究に基づくものである。検討できるのは、仮説検定（例: キャリブレーションスロープ＜１、事前に規定されたc-indexの減少）、もしくは、より好ましいのは新たなデータで測定された精度と正確度についてフォーカスすることである。

多くのシステマティック・レビューで、開発と検証両方の予測モデル研究について、サンプルサイズについての理論的根拠や過剰適合についての言及がなされてないことが観察されている(34, 54, 255)。

## 欠測値

### Item 9. 補完方法の詳細を含めて、欠測値がどのように扱われたか記述する（例: コンプリートケース、単一補完法、多重補完法）。[D; V]

**例**

我々は臨床的な変数と冠動脈造影CTの結果に依存してランダムに欠測が起こっていると考え、多重補完法を用いた。欠測値はアウトカムとともに冠動脈造影CTの結果を考慮した上で、他のすべての予測因子を基に推測された。我々は同一の既知の情報に欠測補完に伴う不確実性を反映する差異を加え、２０のデータセットを作成した。667(2%)の臨床データ項目が代入された。我々の研究ではカテーテルによる冠動脈造影を行われた患者は少なかった。カテーテルによる冠動脈造影を行われた患者に限定した解析では、確証バイアスの影響を受ける可能性がある。そのため、我々は全ての他の予測因子に加えて、冠動脈造影CTの結果を補助変数として用いて、カテーテルによる冠動脈造影のデータとして代入した。2つの処置の結果はよく相関しており、特に冠動脈造影CTの結果が陰性の場合顕著であった。この強い相関は、両方の処置を実施された1609人においても認められていた(ピアソンの相関係数 r=0.72)。それゆえ冠動脈造影CTの結果は予測モデルの予測因子としては含まれなかった。我々のアプローチは、カテーテルによる冠動脈造影が行われなかったとき（これは感度分析として調べられた）冠動脈造影CTをアウトカム変数として用いることに似ている。しかしながら、他の予測因子と欠測値の不確実性を同じく考慮に入れていることから、このアプローチはより洗練されている。我々はカテーテルに基づく冠動脈造影について、3615 (64 %)のアウトカム変数を補完した。多重補完法は、Stata/SE 11(StataCorp) (256)を用いて実施した。[診断; 開発]

もし、アウトカムが欠測していた場合、患者データは解析から除外した。多重補完法をアウトカム以外のデータの欠測に対して用いた。SAS callable IVEware (Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI)を用いた。多重補完法はアウトカムのない患者を除外することによって起こるバイアスを最小化して欠測値を扱う妥当で効果的な方法だと示されている。加えて多重補完法は欠測値の割合が大きくなったとしても妥当な方法である。多重補完モデルに含まれる変数は、4つのアウトカム、年齢、性別、ICD-9 Eコード、救急外来でのGlasgow coma score、病院外でのGlasgow coma score、Injury Severity Score、外傷のメカニズム、外傷チームへの届け出であった。多重補完法ので、10の補完されたデータセットを作成した。ROC(receiver operating characteristic)曲線の曲線下面積を標準的な方法で10の補完されたデータセットごとに統合した。補完されたデータセットからROC曲線の曲線下面積を統合するにあたってよく採用されている方法はないが、我々はそれぞれの感度と特異度を10の補完されたデータセットから平均しプロットすることで、結果の曲線を作成した(257)。[予後; 検証]

我々はデータを開発（トレーニング）と検証（テスト）に分けた。開発のデータには最初の５年間の全ての手術が含まれた。検証のデータには残りが含まれた。データの信頼性を確保するため、鍵となる予測因子（年齢、性別、術式、埋め込まれた弁の数と場所）が欠測している患者については除外した。加えて、開発のデータからは、残る予測因子が3つより多く欠測している患者を除外した。モデル作成過程では、半数未満の患者でしか測定されていない予測因子については、使用しなかった、その結果左室拡張末期圧、肺動脈楔入圧、大動脈弁圧差、活動性の心内膜炎が除外された。検証のデータからは、リスクモデルに使用されている予測因子が欠測している患者については除外された。欠測の結果から除外された患者がバイアスを生んでいるか調べるため、我々は鍵となる術前の患者特性を組み入れられた患者と比較した。開発データのうち、残る欠測がある変数については、多重補完法を用いて補完した。5つの異なるデータセットが作成された（258）。[予後; 開発; 検証]

**解説**

ほとんど全ての予測研究はアウトカムや予測因子の欠測を持っている。しかし、欠測について明示して議論している研究はほとんどなく、欠測について統計的に考察している研究は更に少ない(34, 45, 53, 259)。欠測について言及しない場合は、欠測値のある対象者はどの解析でも除外されていると考えることが自然である、このことはいわゆるコンプリートケースアナリシスにつながる。データが完全な研究対象者のみを組み入れることは、効率が悪い（このことはサンプルサイズを大きく減らすだろう）だけでなく、データの欠測のない対象者が研究全体の代表性を持っていないときには、ゆがんだ結果をもたらす。(つまり、こういった対象者は選択的なサブサンプルである)(**Box D**)。予測モデルの開発または検証を行う研究において、この選択バイアスがあると、全ての分析可能であったはずのデータセットが得られた場合とくらべて、予測因子とアウトカムの関連（モデルの開発）、とモデルの予測能（モデルの開発と検証）に関する推定が異なった（ゆがんだ）結果となる。いまやほとんどのよく用いられている統計パッケージ(Stata, R, SAS)に組み込まれている多重補完法はすべての欠測に対する推定（補完）と、つづく多重補完されたデータセットに対する分析を可能とする。我々は、予測モデル研究で多重補完を行った後で、関心のある推定（回帰係数、予測能の評価）を統合するための既存のガイダンスを読者に示す(**Box D**) 。

予測モデル研究の著者は欠測値についての詳細を注意深く報告し(item 13b)、どのように扱ったか記述する(item 9)ことが推奨される。もし分析に際して欠測値がある対象者が除外されている場合、このことは組み入れ基準に、除外した論理的根拠とともにはっきりと言及されるべきである(item 5b)。

既存のガイダンス(56, 200, 259)に基づいて、欠測値がどのように扱われたか報告する際に含まれる重要な項目は、Table 4に示されている。予測モデルの開発と検証両方を行う研究について、著者は両方のデータセットについて、どのように欠測値が扱われたか全ての差異とともに記述するべきである。

予測モデルの開発と検証研究について、採用された方法論とその報告について評価しているシステマティック・レビューでは、欠測値についての報告とどのように扱われたかが不十分であることで一致している(34, 43, 45, 53, 56, 59, 60, 62, 64, 66, 70, 71, 76, 78– 84, 88, 93, 122, 176, 260, 261)。

#### Box D. 欠測値

予測因子もしくはアウトカムについての欠測は、診断モデルおよび予測モデルの開発と検証を含む全てのタイプの医学研究において起こりうる。一方で、ほとんどの統計パッケージは、どのようなデータを分析する際も欠測値のある対象者を除外する。結果として生じるいわゆる”可能なケース{available case}”もしくは”コンプリートケース{complete case}”分析は欠測値を扱う際に最もよく行われるアプローチである。しかしながら、研究で用いられたそれぞれの変数の欠測値が少数であっても、多変量解析では多くの患者が除外されることになる。もし、除外された対象がもとの研究サンプルから完全にランダムに選ばれていたとすれば、欠測値がある対象を単に除外することは、結果の信頼性には必ずしも影響しない(195-200)。しかしながら、欠測値のある対象者が元の研究対象者からランダムに発生していなかった場合、コンプリートケース分析の結果はゆがむことなる。バイアスの程度は欠測値のある対象者の数を含む、いろんな要素に依存する(10, 195–201, 492)。欠測値についてカテゴリー分けを用いることはゆがんだ結果を示すとされてきており、予測モデル研究では、臨床的に理にかなっておらず避けるべきとされている(195, 196)。

欠測が起こる可能性が、研究において観察された変数、予測因子、アウトカムのいずれにも関連がない場合、欠測データは”完全にランダムな欠測{missing completely at random}” (MCAR)と記述される。他の観察された変数に関連してある変数の欠測が起こっている場合は、 "ランダムな欠測{missing at random}" (MAR) と記述される。欠測値そのものを含む観察されていない変数に関連して欠測が起こっている場合は、 "ランダムではない欠測{missing not at random}" (MNAR)である(493, 494)。欠測がMCARで起こっているか、観察された変数に関連しているかを判断するためにデータを検証することは可能であるが、実際にMARであること、ましてやMNARであることを証明することは一般には不可能である。

単に欠測のある対象者を全て除外したり、欠測インディケーター法を用いるのとはちがい、観察された変数に関連した欠測、つまりMARを仮定する、より効果的な方法がいわゆる補完法である。こういった補完法には、全体の平均や中央値を補完すること、層別化、サブグループ化した補完、多変量モデルを用いるものがある。後者の補完アプローチには一度で行えるもの(単一補完法)や一回以上行うもの（多重補完法）がある(493-495)。

多重補完法はより好ましい方法として推奨されており、より正しい標準誤差とP値を導く。単一補完法では、標準誤差やP値は過少に評価され、誤った偶然の発見(訳注：アルファエラー)を増やすとされる(103, 195–200, 492, 496)。多重補完法には、観察されたデータをもとに予測された分布から抽出された欠測値が補完された複数のデータセットを作成することが含まれる（493, 497）。多重補完法についての標準的な教科書には、5から10のデータセットで十分だとされている。しかしながら、最近になって、欠測の情報の割合に関連して、より多く補完を行うことが提案されている(495)。最終的に、一方で補完された変数についての不確実性を考慮に入れながら(196–201, 492, 495, 499, 500)、標準的な統計分析をそれぞれの補完されたデータセットについて適応することができる、そうして、それぞれの回帰係数やモデルの予測能を全体で推定して統合することができる（Rubinの方法[494]）(item 10d)(2,498)。

## 統計解析手法

### Item 10a. 予測因子が解析に際してどのように扱われたか記述する。[D]

**例**

年齢、血糖値、ヘモグロビンなどの連続変数は、制限3次スプライン{restricted cubic splines}回帰モデルを用いて非線形性を評価した際、アウトカムに対する良い線形性の関係を認めた。（262）[予後]

連続変数である年齢、BMI（body mass index）および期間(年数)の、アウトカムに対する非線形性を証明するために、区分多項式{Fractional polynomials}を用いた。（258）[予後]

これらの予測変数と肺がんのリスクにおける非線形性を、制限3次スプライン回帰モデルを用いて推定した。年齢、喫煙指数（パック・イヤー）、禁煙期間および喫煙期間は、喫煙者におけるこれらの変数のパーセンタイル分布に基づく結び目留置{knot placement}を用いて調整した。年齢の結び目は55, 60, 64, 72歳であった。喫煙指数（パック・イヤー）は3.25, 23, 25, 63パック・イヤーであった。禁煙期間は0, 15, 35年であった。喫煙期間は、8, 28, 45年であった。（263）[予後]

**解説**

多くの予測因子が、連続変数として測定される。研究者は、解析においてこれらをどのように扱うか決めなければならない。臨床予測モデルを作成する研究において、連続変数のカテゴリー変数への変換はよくみられるが、この方法には大きな懸念がある。**Box E**は、連続変数は連続変数として扱うべきか理由、および、予測変数とアウトカムとの関係が線形または非線形かを評価しなければならない理由、について説明している。

先行する臨床的なコンセンサスがない場合、連続変数をカテゴリー化または二分する際にはデータに基づかない方法{a non-data-driven method}が推奨される。

p値を最小にするために最適なカットオフ値を選択することは、例外なく避けるべきである。(264, 265)このようなデータに基づく方法は、モデルの過剰適合そしてモデルの楽観的な性能を導き、予測変数とアウトカムとの楽観的なもしくは誤った関連を導くことになるため、大きな欠陥となる。

カテゴリー変数については、データを解析する前に、操作をしたほうがよい。とりわけ、まれなカテゴリーはなくしてグループ化したほうがよいであろう。例えば、非常にまれな組織型は、ひとつの「その他の組織型」としてまとめたほうがよい。全てのカテゴリー化の方法は説明すべきである。(item 10bも参照)

著者は、全ての解析においてどのようにそれぞれの予測因子を扱ったか明確に報告すべきである。

特に、すべての連続変数のカテゴリー化における、カットオフ値やその選択方法を含む、方法論的または臨床的な意義を報告すべきである。また、連続変数のまま扱う場合は、連続変数をオリジナルスケールのまま使うか、対数変換のように他のスケールに変換して使うか明確にすべきである。また、個々の予測因子を線形、非線形のどちらにモデル化して使用するのか報告すべきである。非線形の場合は、その方法を具体的に（区分多項式、制限三次スプライン等）詳述すべきである。もし、予測因子を線形として扱う場合は、アウトカムとの線形性の仮定を確認したかどうかを報告することが望ましい。過度なレバレッジ効果{leverage effects} (2, 266)を防止するために、極端な値は、極端さを減らす方向に変換した方がよいであろう。

研究者は、このような極端な外れ値を削除した場合などは、異常値を修正または削除したかどうかを報告すべきである。

解析における予測因子の処理方法についての情報は、方法の一部として示すが、結果の表にあるカテゴリーの定義を示すことにも役立つ。(items 13b と13c).

すでに発表された臨床予測モデルに関するレビューでは、一般的に連続変数の予測因子をカテゴリー化しており、その多くは二値変数である。(34, 41, 43, 45, 53, 54, 62, 63, 267, 268)

動脈瘤性くも膜下出血に関する11の研究のレビューでは、すべてのモデルにおいて年齢は、二値変数として扱われていた。癌に関する予測モデルのレビューでは、その30%（12/45）が、最終モデルで使用したすべての予測因子の明確なコーディングを提示していなかった。その他のレビューでも、解析において連続変数の予測因子がどのように扱われたか、しばしばはっきりしなかった。

#### Table 4.　欠測値についての報告に関するキー・インフォーメーション

**方法の中で:**

・予測因子およびアウトカムの欠測値の扱い方についての明確な記載（例: コンプリートケース、単一補完法、多重補完法）

・すべての欠測についての考えられる理由

・補完法に基づく解析（単一補完、多重補完）:

・使用した詳細なソフトウェアについての紹介（特定の補完法を含む。例: ICE, MICE, PROC MI, Amelia, aregImpute）

・予測変数の補完にアウトカムを用いたか、アウトカムの補完に予測変数を用いたかなど、補完作業で使用した変数の表示

・補完モデルにおいて、予測変数、二値変数、カテゴリー変数がどのように扱われたかの説明

・補完モデルに相互作用が含まれていたかの明示

・多重補完を用いた場合の補完回数についての報告

**結果の中で:**

・個々の欠測数（すべての欠測値、欠測値1、欠測値2）

・欠測値の数（予測因子およびアウトカムごとに）

・補完前のデータと完全に補完したデータの比較。これによって、研究における特定の変数（予測因子やアウトカム）に生じた欠測が、完全にランダムな欠測か、観察された測定項目に関連する欠測かの傾向を示すことができる(Box D)。

#### Box E. 連続変数の予測因子\*

予測因子の多くは、連続変数として記録されるものの、しかし1つもしくはそれ以上のカットオフを用いてカテゴリー変数に変換される(item 10a)(501)。共通する理由は解析をシンプルにすること; 予測因子とアウトカムの相関についてはしばしば知られていないため、予測因子や予測モデルを臨床家が使いやすいようにすること; 絵で見せやすくすること（例、カプランマイヤー曲線）である。予測モデルによって推定された確率をカテゴリー化することは決断をするのに必要である、連続変数である予測因子をカテゴリー化してモデルに入れることは統計解析には必須ではないということを認識しておくことは重要である。シンプルにした解析の見かけ上の優位性は高く付くことを以下に説明する。

***カテゴリー化***

カテゴリー化によって研究者は予測因子とアウトカムの間にある相関についての強い前提を避けることができる。しかしながら、このことは情報を捨てることで成り立っている。予測因子を二値変数化(2つのカテゴリー)したときに情報の損失は最も明らかである。予測因子に対して異なるカットポイントを使うと結果が変わり得ることはよく知られている（例、モデルの予測能）。しかしながら、もしカットポイントがデータの多変量解析に基づいて、特に最も小さいP値になるよう選ばれた場合は、予測因子のP値は小さくなりすぎて、予測モデルの予測能は楽観的過ぎる結果になってしまう(264)。

事前に規定されたカットポイントであったとしても、二値化は統計的に不十分であり、するべきではない(265, 502–505)。なによりも、カットポイントが人々を異なるリスクグループに分類する助けとなるために必要であるなら、モデルの予測した確率やリスクに基いてカットポイントは選ばれるべきである(30,265)。

連続変数を3またはそれ以上にカテゴリー化することで情報の損失は減るが、臨床研究ではほとんど行われていない。もしある変数をカテゴリー化して、階段的なモデル化を行っても、円滑な関連を記述するには不十分である(266)。

***連続変数のままにする***

線形関数の相関は、予測因子の連続性を保つのに最もよく行われている。これはしばしば受け入れられる仮定ではある。しかしそれは誤りである可能性があり、その場合は、重要な予測因子が含まれなかったり、推定された予測因子とアウトカムの関係が知られていない「真の」関係から実質的に異なっていたりするような、誤って作成されたモデルの原因となる。線形性の確認は、非線形性の型を許容した場合に当てはまりがどこまで改善可能かを調べることによって、行うことができる。長い間、二次もしくは三次多項式が非線形相関をモデル化するのに用いられてきた、しかしより一般的な区分的多項式(FP)関数の一群によって、しばしばあてはまりの改善をもたらす、多くの単純な関数が作られた(506)。FPの詳細およびモデルを選択することは、単純で理解可能な結果の表示と同時に行うことができる(266, 297)。

スプライン関数、特に三次スプラインに制限されたもの、は連続予測因子の関数関係を調べるもう一つの方法である(112)。制限三次スプラインは、標準的な三次スプライン関数よりも推奨されている。その理由は分布の裾を線形にすることで、予測因子の分布の裾でより当てはまりやすいからである(112, 507)。スプライン関数は非常にフレキシブルである、しかし同時に予測因子を選択し、分布関数を作るために、広く受け入れられている方法はまだない。それ以上に、単変量のスプラインモデルであっても、回帰係数を表示することはしばしば非常に複雑であるため、報告は予測因子とアウトカムの関係をプロットすることに制限される。

\* このboxのテキストについてはBox 4のリファレンス108と実質的に同じである。

### Item 10b. モデルのタイプ、モデルの作成過程（全ての予測因子の選択過程）、内的検証の方法を明記する。[D]

予測モデルを作成するのに使われた全ての統計方法は報告されるべきである。一般原則として、知識のある読者がオリジナルデータにアクセスした際に、報告された結果を検証できるくらいに十分な詳細が与えられるべきである([www.icmje.org](http://www.icmje.org))。さらに、読者がある統計方法が選択された理由を理解できるようにすべきである。

予測モデルを作成する際に従うべき多くの解析戦略がある。選択は解析のそれぞれのステップで行うことになる(2, 112, 266, 269)。モデル化戦略に際して、決めるべきことの中には、医学的な文脈と同様にデータの情報を必要とするものがある。例えば、臨床的な応用可能性を増やすために少数の重要な予測因子のみ使ったモデルを作成しようと考えた場合(items 3a, 19b, and 20)、予測モデルの性能を犠牲にすることになる。

多くの予測モデル研究で重大な問題としては、多くの異なる解析が実施されていても、最も良い予測モデル（識別能が最も良い）しか報告されないことである（1）。こういったデータに基づくモデル選択では過剰適合して、楽観的な性能を持つモデルを選択することになる。この過剰適合は、同じ対象からの新たなデータでモデルが評価されたときに明らかになる(270)。それゆえ、研究者は実施された解析の範囲について包括的な視点で提示する必要がある。もし必要であれば、解析を行うのに使ったコンピューターコードも含めた統計解析の全てを明記して補足として与えられるようにする(item 21)。理想的には、このコードは対象者個人のデータと一緒になって、完全な再現性を持つべきであるが、このことはデータへのオープンアクセスが合意されていないと実現可能性は乏しいであろう(271)。

以下の章で、我々はいくつかの章立てのもとモデルの開発分析に特異的な側面について考える。全ての側面は、いくつかの研究では重要ではないだろう。二値と生存時間アウトカムについてのより広範な統計解析方法についての議論は他稿を参照されたい(2, 12, 112, 266, 272–277)。

#### モデルの型

*例*

我々はコックス比例ハザードモデルを作成用データセットで用いて男女別に心血管疾患の診断が初めて記録されるまでの期間と、それぞれの潜在的なリスク因子[予測因子]との相関について係数を推定した(278)。[予後]

全ての臨床、検査予測因子を多変量ロジスティック回帰モデルに組み込んだ（アウトカム: 細菌性肺炎）(279)。[診断]

*解説*

様々な種類のモデルが医学予測研究には用いられる(112)。ほとんどのモデルは多変量回帰分析からの派生である。ロジスティック回帰モデルは、疾患モデルでの疾患のありなし、短期的な予後イベント（例、30日以内の死亡）といった二値アウトカムに対して最もよく適用される。パラメトリックモデルを用いることも可能であるが、セミパラメトリックのコックス比例ハザード回帰モデルは長期の予後アウトカム（例、10年心血管疾患リスク）の生存時間アウトカムにもっともよく用いられる(280, 281)。

著者は使われた回帰モデルをはっきりわかるようにするべきである。もしロジスティック回帰分析が長期予後アウトカムを予測する生存時間アプローチに対して使われた場合、はっきりと理論的な根拠が報告されるべきである。ロジスティック回帰モデルを用いた長期アウトカム予測モデルを開発（検証）する場合、もともと全ての対象者が全期間にわたってフォローされていることが必要になる。回帰モデルは、二値、多項、順序、連続、またその他の変数に対して使うことができる(2)。他の予測モデルには回帰ツリーやニューラルネットワークとサポートベクターマシンといった機械学習法がある(275)。こういった他の手法が用いられた場合、その選択の動機を提示することが勧められる。

#### モデル作成前の予測因子選択

*例*

我々は 先行するメタアナリシスとレビューの結果に基づいて以下の視点からリスク因子を選択した。プライマリ・ケアセッティングでの使用しやすさ ; 習慣変容や治療介入により リスク因子が変可能なこと ( 例、喫煙)。 しかし、我々はEPISEMによって構成された二つのコホートで既に用いられている因子のみを用いることになった(282)。[予後]

候補となる変数には、それぞれのデータソースから全ての人口統計属性、疾患関連の因子、ケアのパターン、といった、既報で集中治療を受けた後の死亡に関るリスク因子であることが示されているが含まれた。変数は当初文献のレビューと集中治療医、総合医、集中治療でトレーニングを受けた看護師、疫学者、統計家を含むエキスパートグループのコンセンサスから選択された。確認されたセットは、ANZICS APDに慣れている5人の集中治療医と生物統計家によってレビューされ、承認された。(283)。[予後]

臨床的な重要性と我々のコホート（症候性の心房細動で救急外来を受診した患者）での組み入れ時の記述統計の結果に従って、多くの変数セットから、我々は12の予測変数を予測ルールに組み入れた。特に我々は30日間に有害事象を経験した患者と経験しなかった患者について、組み入れ時の特徴をレビューし、2群で明らかに違った予測因子として表現されたもの、臨床的重要性と合理性から、50の候補となる予後因子から12の予後因子をモデルに組み入れた。…共線性を制限し、よりシンプルなモデルにするため、スピアマンの相関係数を、12の予測因子の中で臨床的に合理性がある関連の間で計算した。特に、スピアマンの相関係数は以下の臨床的に合理性のある関連について計算した: (1) 高血圧の既往とベータブロッカーと利尿剤の使用、(2) 心不全の既往とベータブロッカーの自宅使用、利尿剤の自宅使用、身体診察での末梢浮腫、救急外来での呼吸困難(284)。[予後]

*解説*

しばしば、研究者が最終予測モデルに入れようと望んでいるよりもより多くの予測因子が使用可能である。そのため、いくつかの予測因子の選択方法が必要となり、それぞれに強みと弱みがある (**Figure 2**)。

候補となり得る多くの予測因子を減らすのに疑いの余地のない方法は、事前にどれを除外するかを決めておくことである(item 7a)。ここで外的なエビデンスを探すことになるだろう、例えば重要な文献の批判的考察を理想的には正式なシステマティック・レビューの形をとって行う。医療の専門家の知識も、候補の予測因子の数を減らすのに重要である。

実際の解析を行う前に予測因子を除外する場合の考慮点として、予測因子の測定信頼性が低い場合、測定のための費用や対象者の負担が大きい場合などが挙げられる。後者の場合、そういった予測因子を含むものと含まないもので、ますます複雑性を増したモデルを作成することがある(262)。また、密接に関連した予測因子については、動脈硬化の症状があること(285)のように合計スコアとして合算させることが可能である（例、正式な統計クラスター化や主成分分析の手法を用いる）、また共線性がある2つの予測因子のうち1つを、事前に選択するために予測因子の間の相関を推定することもできる(例、相関係数を使う)。

#### 3. モデル生成における予測因子の選択

*例*

変数減増法を用いた多変量ロジスティック回帰分析においてP値が0.05より大きい変数を除外したが、臨床的意義の大きい変数[予測因子]については含めるようにした。また、臨床に用いられるガイドラインに記載のあるリスク因子[予測因子]についても追加で評価を行った（286）。[診断]

*解説*

上記のように予測因子の事前選択を行っても、予測モデルに含めたい項目数より多い予測因子が残ることがある(Figure 2)。その場合には、各予測因子の重要度によって更なる選択を行うことは可能である、または残った予測因子全てを用いたモデルに当てはめる（287）。

予測因子の選択の一つの方法としては、予測したいアウトカムと個々の予測因子の未調整の相関係数（単変量解析）で評価し、関連の強いものをモデルに投入するという方法、すなわち多変量解析を行う前に事前選択することが挙げられる。単変量解析でアウトカムとの有意な相関がない予測因子は、限定的な予測能しかないため予測モデルから省くことができるだろうというのが理由である。この方法はよく行われているが、重要な予測因子がデータセットのニュアンスによっては除外されたり、他の予測因子の交絡の影響によって除外されたりという可能性があり、推奨されない（2, 112, 235）。このように、(単変量解析で)有意でないことは予測因子として重要でないということを必ずしも意味しない。しかしながら、もし単変量解析を行ったのであれば、選択基準（例、有意水準）、各単変量解析におけるサンプルサイズ（アウトカム発生数を含む）を含めて記載すべきである。(items 13b 、14b、**Figure 2**).

一般的に用いられる方法としては、多変量解析のモデルで、自動的に変数を選択するものである。変数増加法、変数減少法、両者の組み合わせなどが最新の統計ソフトで利用可能である。変数減少法では、全ての変数を最初にモデルに投入し、事前に設定された条件（P値、赤池情報量基準[AIC]など）を満たすまで変数を省いていく。変数増加法では、事前に規定した条件を満たすまで、予測変数を１つずつ加えていく。

自動変数選択に際しては、変数減少法が一般的に好まれる。それは、モデル生成の過程で、全ての変数間の関連が検討されるからである。サンプルサイズが小さい場合に、自動変数選択法では、過剰適合した、楽観的なモデルが得られることがある(2, 23–25, 32, 112, 289, 290)。しかしながら、この予測変数の選択の過程で、過剰適合した程度は、いわゆる内的検証の過程で推定し、説明することができる(**Box C**、**Figure 1**)。

自動変数選択において重要なのは、変数が採用される基準である(2)。しばしば、予測因子の有意水準は、一般的な仮説検定のように0.05に設定される。しかしながら、シミュレーション研究では、特に、サンプル数が少ない場合において、より高い値を採用すべきであることが示唆されている(25)。そのような場合、変数選択の基準にAICを用いることは魅力的な選択肢である; AICは変数の数を抑えつつ、有意水準0.157で適合度の良いモデル選択を可能にする(2, 112, 291, 292)。

多変量予測モデルのシステマティック・レビューで、予測モデルの作成方法が不明確であることがよくあると指摘している(34, 43, 54, 81, 182)。例えば、動脈瘤破裂によるくも膜下出血に関する11の予測モデルの36%が変数選択の過程が不明確であったという。

#### 4. 交互作用項

*例*

臨床的に意義のある交互作用のある項目がモデルに採用された。その統計学的有意性については、タイプⅠエラーの増大を避けるためにグループとして評価した。交互作用のある項目は全てグループとして除外し、結果が有意でない場合には、モデルに再度組み入れた。例えば、自宅でのβ遮断薬と利尿薬の使用の関係、浮腫の所見と心不全の既往の関係を評価した(284)。[予後]

*解説*

ほとんどの予測モデルは、それぞれの予測因子の効果が相加的であると仮定している。ここで言う相加的とはモデル作成の際の点数について仮定していることである（ロジスティック回帰分析の対数オッズ、コックス回帰分析の対数ハザード）。ここで言う相加的であることは、元のオッズやハザードスケールそれぞれに対しては、相乗的であることを意味する(273)。この仮定は、各予測因子の予測効果が、他の予測因子の影響によらず一定であることを意味する。この仮定は、予測因子間の交互作用を検定することで統計学的に評価することができる(112)。これまでに報告された予測モデルでは、予測因子間の交互作用を考慮したものはほとんどなく、そのことから交互作用について評価している研究者もほとんどいない。交互作用項が予測モデルの予測性能を向上させることがほとんどないことを考えると、このアプローチはおおむね理にかなっている。

もし多くの交互作用が評価され、最も影響の強いものだけが予測モデルに投入されると、モデルの過剰適合が起こり、非常に楽観的なモデルとなる(2)。特に、サンプルサイズが小さい場合、予測因子間の交互作用の評価については、全ての因子について評価するのではなく、理論的根拠に基づいて事前に定めたものに限るべきである。代替案としては、サブグループごとに異なるモデルを開発することである: 例、男性と女性、成人と小児(278)。サンプルサイズの減少とモデルの過剰適合の恐れがあるため、この方法が採用されることは稀であり、サンプルサイズが大きい場合にのみ考慮されるべきである。

生存分析モデルでは、しばしば予測因子の影響が時間によらず一定(つまり、ハザードが比例性)であると仮定することもある。これは時間による交互作用がないと仮定することと同じである。比例ハザードモデルの仮定を検定することを推奨する意見がある一方で、統計学的に有意な非比例効果で調整した場合に、変数選択の過程と同様に過剰適合や楽観的なモデルとなることを警告する意見もある(2, 112)。

著者は、もし実施したのであれば、相互作用および生存分析モデルにおけるハザード比の比例性についてどのように評価したのか報告するべきである。

#### 5. 内的検証

*例*

予測モデルの内的検証は、ブートストラップ法を用いて両方のモデルについて、組み入れ患者と似ている将来の患者での性能を実際的に見積もることで評価した。変数選択を含むモデル形成の過程を…元のサンプルから200例を抽出することで繰り返した。選択した予測モデルの性能と、各ブートストラップサンプルから開発した予測モデルの性能を元のサンプルで評価した。性能の評価は、各ブートストラップ法のサンプルにおいて両方のアウトカムに関するROC曲線下面積、感度、特異度、および神経手術実施に対する感度を100%とした場合のCT施行頻度の減少によって行った(286)。[診断]

*解説*

モデルを生成した際に用いたデータにおけるモデルの予測能は見かけの性能{apparent performance}と言われている(12, 293, 294)。特に、小さなデータセットから予測因子の選択を行った場合において、多くの予測モデルは過剰適合し、見かけの性能は楽観的な数値を示す(23–25, 32, 290, 295)。予測モデルの性能の評価は、横断検証{cross-validation}やブートストラップ法といった再サンプリングを通して行う方が良い、これらは全て内的検証と呼ばれる(**Figure 1**、**Box F**)(12)。我々は全ての予測モデル開発研究において、特に、外的検証が行われない場合には、何らかの形で内的検証を実施することを推奨する。予測因子の選択が単変量および多変量解析での予測能の強さやp値によってなされる場合、モデルの構造に不確実性が含まれる(292, 296)。横断検証と比較したブートストラップ法の強みは、モデル作成の過程における予測因子の選択の効果、およびモデルの過剰適合、見かけの性能が、繰り返し抽出された各サンプルにおける予測因子の選択過程で定量化されることである(292, 296–298)。更に、ブートストラップ法では、いわゆる調整または補正係数を推定することで、予測モデル(つまり回帰係数)とその性能評価(item 16)のシュリンケージ、そして過剰適合を調整できる(**Box F**)。

予測因子の選択、（訳注: 対数などへの）変換、他の変数や時間との交互作用の検定を含む、モデルの適合の全ての過程をそれぞれのランダム、またはブートストラップ法で得られたサンプルにおいて評価することは非常に重要である。臨床研究においてこのプロセスが省かれていることはよくあるが、妥当性検証のためのサンプルにおいてさえ、適合度の評価にバイアスをもたらす(299, 300)。ブートストラップ法で得られたそれぞれのサンプルにおいて同じ予測因子を用いたモデルの適合度を検証するのは、予測モデルが全ての予測因子を採用していない限り（いわゆる全モデルアプローチ{full model approach}）妥当な方法ではない。著者は、全ての内的検証についての詳細を報告する必要がある。

過剰適合、楽観的な評価、較正の誤りは、シュリンケージまたは罰則{penalization}法によって対応することができる(287, 290, 294, 301)。モデルが稀な事象に対するものであったり、予測因子の数が非常に多い、サンプルサイズが小さいなどの場合においては、投げ縄{lasso}法とその変法が一部では人気である (24, 302, 303)。しかしながら、予測因子が少ない場合においての有用性については明らかではない(291)。もしこういった方法を用いた場合は、その詳細を報告すべきである。(例、投げ縄、リッジ回帰、発見的{heuristic}シュリンケージなど)。

一般内科雑誌で出版された予測モデルのレビュー論文において、内的検証は14論文中5件でしか行われていなかったという報告がある(34)、他のレビューでも同様の報告がなされている(43, 53, 55, 64, 66, 71, 75, 76, 88, 93–95, 304, 305)。

#### Box F. 内的検証

予測モデルの開発において、その見かけの性能を楽観的にさせうるいくつかの要因がある。これらの要因には、アウトカム発生数に対して予測因子候補が多い（有効なサンプルの不足）、予測因子の選択方法（有効なサンプルの不足と合わせて起こりやすい）、連続変数のカテゴリー化などが含まれる(2, 12, 23–25, 32, 112, 290)。よって、開発データでのモデル性能のより正確な推定が重要である。これはブートストラップ法、横断検証などの再標本化法によって、いわゆる"内的検証"を行うことでなされる。

**見かけの性能**

予測モデルの見かけの性能とは、そのモデルの開発に使用したデータセットから直接見積もられるモデルの予測能のことである。予測モデルは、開発に使用したデータセットに適合しており、小さいサンプルでは、楽観的な性能（ゆがんでいるが安定している）を示してしまう。しかしながら、この過大なな見積もりの程度は、サンプルが大きくなることによって減少する(32)。

**サンプル分割検証{Split-sample validation}（”データの分割”）**

古典的なデータの分割による内的検証においては、利用可能な開発データセットを２つに分割する。1つはモデル開発用、もう1つは妥当性の検証用である(**Figure 1**、**Box C**を参照)。典型的には、2つのデータセットは元のデータからランダムに生成される(例、50 :50または70: 30)。この方法は予測モデル研究でよく行われているが、いくつかの欠点がある。まず、全てのデータをモデル開発に使用できないため、非効率的である。また分割された2つのデータセットは偶然によってしか変わらないので非常によく似ている。（モデルの検証において、開発データで示されたものと同様の性能を示す。）そして、比較的小さいデータセットでは分割の仕方によって異なる結果が得られる可能性がある(23, 25, 32, 295, 508)。モデル開発をするためにどれくらいのデータが使われるべきか、評価をするためにそれぞれにどのくらいのデータを確保するべきかについてはよく分かっていない(item 8参照)。この方法を理にかなったものにするためには大きなサンプルサイズが必要となる、そうすれば見かけの性能はモデルの性能を適切に推定しているだろう(2, 32)。サンプルサイズが十分に大きい場合には、他のより良い方法として時間(temporal validation）または研究の実施された場所(geographic validation)によってデータセットを分割する方法もある(19, 20, 26)。

**横断検証{Cross-validation}**

横断検証は、サンプル分割検証のバイアスと性能推定の変動を減らす拡張法である(32)。例えば、10組横断検証では、データをランダムに10の同じサンプル数のグループに分割する。予測モデルは10のうち9グループを用いて開発され、残りの1グループでその性能を検証する。この全ての過程を10回繰り返し、各グループが検証用に用いられる。モデルの性能は、10回の反復の平均となる。

**ブートストラップ法**

ブートストラップ法では、モデル開発に全てのデータを使用することができるだけでなく、また過剰適合やモデル開発の不確実性に対処できる、それゆえ最終的に得られたモデルにおける過大なな評価を定量的に評価することができる。また回帰係数の調整および過大な見かけの性能を調整するための、いわゆるシュリンケージ因子を推定することが可能であり、続くモデルの検証および臨床応用においてより良い性能を得ることができる。ブートストラップ法は、下記の過程から成る(2, 12)。

1. 全てのサンプル(n人)を用いて、予測モデルを開発し、見かけの性能を検証する。
2. もとのサンプルからn人の個人を抽出してブートストラップサンプルを生成する(訳注: ランダムに n 人の個人を重複を許して抽出)。
3. ブートストラップサンプルを用いて、モデルを開発する(step 1に含まれる同じモデル作成と予測因子の選択法全てを行う)。
4. ブートストラップサンプルでの見かけの性能(例、c-index)を算出（ブートストラップ性能{bootstrap performance}）
5. ブートストラップサンプルで作成されたモデルのもとのサンプルにおける性能を算出（テスト性能{test performance}）
6. ブートストラップ性能とテスト性能の差から過大な見積もり分を算出。
7. step 2-4を最低100回繰り返す。
8. step 5の過大な見積もりの推定値を平均し、step 1における見かけの性能を引いて、過大な見積もりを調整した性能を算出する。

高次元データ(例、”オミクス”やゲノムワイド関連の研究などの)において、横断検証やブートストラップ法がそれぞれの横断検証、ブートストラップサンプルにおいて全てのステップが繰り返されず、不適切に適用されているというエビデンスがある(299, 509, 510)。これは、過大な性能の推定につながる(299, 511)。またその他のバイアスが重なり、過剰に楽観的な性能が算出されることもある(512)。

### Item 10c. 検証に際しては、予測結果をどのように計算したか記述する。[V]

**例**

前立腺癌のリスク計算の性能の評価のために、各患者の全ての前立腺癌および進行前立腺癌の予測されるリスクをPRC(前立腺癌予防研究　リスク計算ツール)(http: //deb.uthscsa.edu/URORiskCalc/Pages/uroriskcalc.jsp)、SRC[Sunnybrook nomogram–based prostate cancer risk calculator] (www.prostaterisk.ca) を用いて算出した(306)。[診断]

肝臓の脂肪変性のリスクについては、HSI[Hepatic Steatosis Index]をLeeらの計算式を用いて以下のように算出した(307)。[診断]

HS1=

糖尿病あり: 1、なし: 0、ALT: alanine aminotransrerase、AST: aspartate aminotransferase。[診断]

QCancerスコアの算出は、www.qcancer.org/colorectal/ released under the GNU Lesser General Public Licence, version 3で公開されている(308)。[予後]

**解説**

新しいサンプルにおける既存の予測モデルの性能の評価(**Box C、Figure 1**)には、検証用のサンプルにおいて、元のモデル(出版されたものと同様に)による予測を行い、実際のアウトカムと比較することが望ましい(つまり、較正と識別能)(309); item10d。そのため、既存の予測モデルの性能を評価する著者らは、そのモデルの予測結果をどのようにして算出したのか記述することが重要である。予測モデルをそのまま（全ての回帰係数や切片、ある時点でのハザード比などを含む）用い、Web計算機へのリンクを提供し、計算に用いたコンピューターのコードを含めて記載する(item 14)。

複数の予測モデルを報告したり、同じモデルの複数のパターン（たとえば、全ての回帰モデルと単純化されたスコア）を報告している開発研究がある。もし必要であれば、著者らは、どのモデルを評価したのか明記するべきである。予測モデルは、しばしばノモグラムなどのように図示されることがある(item 15b) (310, 311)。ノモグラムは、計算機やコンピューターを用いずに簡単に推定を行うことができるが、大きなサンプルの検証研究には不向きである。著者らは、実際のノモグラムに記入して検証したのか、元となった回帰モデルを用いたのか明確にすべきである。

発表された予測モデルにアクセスしないで、検証、再較正そして改訂することはできない。例えば、10年後の骨粗鬆症または大腿骨骨折のリスクについてのFRAX予測モデル(312)は、世界中のガイドラインに採用されている(35, 37, 313)が、出版されておらず、モデルの独立した評価を行うことは不可能である(314–316)。

既存の予測モデルの検証については、いくつかの誤解がある。まず、検証が、予測因子の選択や回帰係数の推定、（とモデルの性能）といった全ての予測モデル生成のプロセスを再度行い、つづいてこれらの結果を既存のモデル開発研究の結果と比較するという誤解である。もう一つは、検証データにおいて、過去に開発され、出版された最終モデルを、再修正するという誤解である。いずれも結果は、別のモデルを生成することになり、検証を行ったことにはならない(19, 20, 26, 28, 47, 309)。

著者らは、しばしば新たに開発した予測モデルを別のデータセット（例: 後からリクルートされた集団または別の病院でリクルートされた集団）で検証を行うことがあるだろう。開発と検証データでの性能が同じように見えた時、2つのデータセットを結合させて、新たな予測モデルを開発することはまれではない(317)。それ自体は悪いことではないが、このような研究は検証研究とは言わない。検証とモデルの開発ないしは再開発ををこの順番で行ったということになる。新しく得られたモデルは依然として更なる検証を行う必要がある。

### Item 10d. モデルの性能を評価するために使用した全ての指標を明記する、重要であれば複数のモデルを比較する。[D; V]

　予測モデルの予測性能の評価および定量化には数多くの手法が存在する(26, 252, 253, 269) (Box GとH)。ここでは最も広く利用されていて我々が研究者に報告を推奨する方法を挙げる。これらの手法は、より伝統的（統計的）な手法と最近の手法、すなわち様々な程度で、予測することの臨床的な結果を評価する手法、および、ある予測因子を既存の予測因子や確立された予測因子に付け加えたときの予測能を明確に推定する手法や異なるモデルを比較する手法に分けられる。

**1．伝統的な手法**

*例*

我々はTHINコホートに対して、QRISK2-2011リスクスコアの予測能を、較正と識別能について評価した。較正は10年間の予測心血管リスクが、観察された10年間の心血管リスクとどれくらい一致しているかの指標である。これは10分割した予測リスク毎に10分割した対象者を5歳毎の幅で当てはめて、予測心血管リスクと実測心血管リスクの割合を男女別に計算して評価した。リスクスコアの較正は実測の割合と予測割合をプロットし、較正スロープ{calibration slope}を計算することで評価した。

識別能は研究期間中にイベントを経験する人としない人を区別するリスクスコアの性能である。この評価法は受信者操作特性曲線{receiver operating characteristic curve: ROC曲線}下面積で定量化され、この値が0.5なら全くの偶然を表し、1なら完璧な識別能を有することになる。我々はD統計量とR2統計量も計算した。それぞれ識別能と説明分散を説明するものであり、打ち切りのある生存データに対応している。D統計量が高ければ高いほど良い識別能を表し、他のリスクスコアより0.1増加することは予後識別の良い指標である(117)。[予後; 検証]

まず第一に、我々は患者の疾患のありなしについて、臨床決断ルールの性能とGP{general practitioner}の判断をROC曲線を用いて比較した。ROC曲線下面積(AUC)が0.5であることは全く識別出来ていないことを示し、AUCが1であることは完璧な識別能を示す。次に、決断ルールの予測確率と実際のアウトカムである急性冠症候群との一致具合を調べるために較正プロット{calibration plot}を描出し、別にGPの判断による予測確率についても同様に較正プロットを作成した。完璧な予測は較正プロットのアウトカムと一致して45°のライン上にある(318)。[診断; 開発]

内的検証を行い、調整したモデルの診断精度を検証データセットで評価した。開発したモデルの回帰式を、検証セットの全てのパン屋職人に当てはめた。予測確率と実測された感作の割合の一致度(較正)はx軸に予測確率をアウトカムである実測された割合をy軸にとってプロットしグラフを使って評価した。予測確率と実測の割合の関連性は切片と傾きによる直線で描出される。切片が0でスロープが1の場合は完璧な較正を意味する。識別能はROC面積を用いて評価した(319)。[診断; 開発]

*解説*

較正と識別能の2つの鍵となる側面が予測モデルの性能を特徴づける。これらは全ての予測モデルに関する論文で報告される必要がある(**Box G**と**H**)。

較正はモデルの予測結果と実測のアウトカムとの一致度を反映する。較正はy軸に実測のリスクを、x軸に予測されたリスクをプロットしてグラフを用いて報告されることが好ましいが、表形式で表現されることもある。

識別能はアウトカムになるイベントのありなしを予測モデルが区別する性能のことを指す。ロジスティックモデルでも生存モデルでも、最も一般的に広く報告されている識別能は、concordance index (cインデックス)であり、これはロジスティック予測モデルのROC曲線下面積に相当する。cインデックスは様々なバージョンが数多く存在する(320)ので、研究者らはどのバージョンを計算したのかを明確に述べるべきである。

識別能や較正に加えて、他にも全体の性能を示す様々な指標、例えば説明分散{explained variation}を示すR2 (321–329) やBrierスコア(330–332)などが報告されている。R2を推定する方法は非常に多くの方法が提唱されてきたので研究者らはどのバージョンを用いたかはっきりと述べる必要がある。生存モデル（例えばCox回帰を用いたもの）では、D統計量が予後を分ける手法として最近提唱された(333)。

同じデータセットで異なるモデルの性能を比較する時にDeLongテストのような統計学的検定を行うことがある(334)。しかし、この検定は比較するモデルの一方が他方を内包（つまり一方のモデルが他方のモデルの全ての予測因子に加えて少なくとも1つ以上の別の予測因子を含む場合）している状況で、同じデータセットでの性能を評価する場合は不適切である。例えば、臨床的な予測因子に新しい分子マーカーを加えたモデルと臨床的な予測因子のみからなるモデルを比較するような場合である(335)。

最後に、新しく開発したどのような予測モデルについても、既存の、出版されたモデルと、理想的には定量的な手法で、比較することが広く推奨される(47, 48)。同じデータセットで複数のモデルの直接比較がなければ、全ての入手可能な予測モデルの中でどれが潜在的により有用かの判断が困難になる。多くのシステマティック・レビューによると、開発された予測モデルや検証される予測モデルについて既存のモデルと同じアウトカムについて性能を比較している研究はほとんどないことが示されている(82)。

較正は主要な特性であり、その評価は広く推奨されている。しかし、多くのシステマティック・レビューによると、多変数予測モデルについて較正はほとんど報告されていないことがわかっている (34, 41, 43, 55, 62, 63, 66, 73–82, 84, 86, 88, 90–92, 94, 122, 176, 180, 267, 336)。例えば、2型糖尿病の予測モデルの研究では較正が評価されていたのは39件中10件(26%)にしか過ぎなかった(45)。識別能は最もよく使われている手法ではあるが、これも必ずしも報告されていない(74, 78, 81, 88, 122, 336, 337)(例えば動脈瘤性クモ膜下出血のモデルの研究では44％だった[81]）。同じデータセットを用いて、モデルの性能を既存の予測モデルと比較している研究は非常に少ない(81, 82, 122)。

#### Box G. 性能評価

リスク予測モデルを開発する際には、その性能を評価するべきである。モデルの性能で最も大事なものは識別能と較正である（item 10dとBox Hを参照）。モデルの開発研究では、主要な関心は識別能にある。というのは、定義上そのモデルの較正は、平均的には良好と考えられるからである。検証研究では、識別能と較正の両者の評価が基本になる(252, 513)。

**較正**はモデルからのアウトカム予測と実測されたアウトカムの一致度を反映する。平たく言うと、例えば100人の対象者について、それぞれが平均の予測リスクがx%だったとして、（診断モデルでは）実際にアウトカムを有する割合がxに近ければ、（予後モデルでは）実際にアウトカムが発生する割合がxに近ければよく較正されたモデルと言える。

較正はグラフを使って報告されることが好ましい。x軸に予測されるアウトカムの確率を、y軸に実測のアウトカムの割合をとる。このプロットは予測リスクの10分割で示されることが多く、全予測確率の幅に渡ってスムーズ(lowess(※訳注:locally weighted scatterplot smoothing))曲線で強調されることが望ましい。これはロジスティック回帰による予測モデル(112, 514)でも、生存モデルでも可能である(515) (item 16を参照)。このプロットは予測範囲でのモデルの誤較正の方向と程度を表し、較正スロープと切片を組み合わせて推定される(515)。スムーズ曲線またはサブグループで示すと、較正のいいモデルは予測結果が較正プロットの45°のライン周囲に示される。完璧な較正はスロープが1で切片が0になる。たっだし、最近ではいくつかの注意事項が指摘されている(516)。

較正プロットはモデルを開発したデータセットではよい較正を示す傾向があり、スムーズ曲線を用いれば完璧な較正になることさえある。較正の切片が0、スロープが1という検定を行うこともある(517, 518)。予測アウトカムと実測アウトカムの確率の比較は表形式（通常予測リスクの10分割）で示すこともできる。

最後に、予測確率と実測確率の一致度の統計学的検定のために Hosmer–Lemeshow検定を行うことが多い。生存モデルについてこれに相当するのは、Nam–D'Agostino検定(519)やGrønnesby–Borgan検定(520)である。このような検定は、較正が良くないことを評価するには、統計学的検出力に限界があり、群分けやサンプルサイズの影響を受けやすい(521–523)、すなわちサンプルサイズが小さければ有意になりにくく、サンプルサイズが大きければほとんど常に有意になってしまう。

その上、これらの検定では較正不良の程度や大きさの情報は全く含まないので、較正プロットが好まれる。さらに、較正（プロット）は、年齢や性別といった鍵となる予測因子との関連性も評価することができる(117, 524)。多項予測モデルについての較正の評価に対するアプローチが最近提唱された(525)。

**識別能**は予測モデルがアウトカムのイベントを発生するかどうかを区別する性能のことを指す。完璧な識別能をもったモデルは、診断モデルの場合はアウトカムを有する全ての対象者、予後モデルの場合はアウトカムが将来発生する対象者の予測リスクが、アウトカムを経験しない全ての対象者の予測リスクよりも高いことを意味する。識別能はいわゆる一致インデックス(cインデックス)で推定されることが多い。cインデックスは、アウトカムを有する対象者と有さない対象者のペアをランダムに選んだ際に、予測モデルがアウトカムを有する対象者により高い可能性を示す確率を反映する(526)。cインデックスは2値のエンドポイントのモデルに対するROC曲線下面積と同じであり、打ち切りを考慮に入れる時間ｰイベント（生存）モデルに一般化することができる。生存モデルについては、様々なcインデックスが提唱されており(527)、研究者らは適切な引用を用いてどの手法を使ったかを明確に述べる必要がある。最近では、2つ以上のアウトカムカテゴリーを有するモデルへのcインデックスの拡張(528)や競合リスク(529)、クラスター化への拡張(170, 171)も提唱されている。

**全体の性能指標**として、説明分散{explained variane}（R2）(321, 324–329) やBrierスコア(330, 331)があり、識別能や較正といった古典的な指標に加えて、直感的には理解しにくい指標だが、報告されることがある。

さらに、R2の推定には様々な方法が提唱されているので、研究者らはどのバージョンを計算し、報告しているのかを明確にするべきである。

**分類手法**には、的中率、感度、特異度のようなものがあり、1つ以上の確率閾値を設定して性能を評価する。こうすることによって、単一の診断検査や予後因子研究で報告されるような正確度(accuracy)や分類手法を推定できる。しかし、そのような二値化とそれに伴う分類手法は情報の喪失につながる。さらに、そのような閾値を設定することは診療に関連があると思わせるが、実際にはそうではないことがしばしばである。

**決断曲線分析(decision curve analysis)**(360, 363–366) は、臨床的な転帰への見通しを提示する。これは、ある閾値においてモデルを使用することの正味の利益を求め、その予測確率閾値における、偽陽性と偽陰性の相対的な価値との関連性を決めることによる。

**純再分類改善度{net reclassification improvement}(NRI)**は、既存のモデルに新しい予測因子を加えることが利益になるかどうかを定量化するためによく用いられるが、2つの別々のモデルを比較するためにも用いられることがある(339, 347, 348, 420, 530)。NRIは正しく再分類される正味のイベントの割合と、正しく再分類される正味のイベント無しの割合を合わせたものである。NRIの上限は連続性NRI（すなわちカテゴリーなし）であり、ここの対象者の予測リスクのどのような変化（増減問わず）も考慮する (347, 530)。

**統合識別能改善度{integrated discrimination improvement} (IDI)**は、診断検査の場合アウトカムを有する人と有さない人の予測確率の差、予後予測の場合アウトカムが発生する人としない人の予測確率の差である(339)。これは2つのモデル（入れ子になっていてもいなくても）について、全ての可能性のある確率閾値に渡って予測の改善または悪化をの程度を推定する。IDIはアウトカムなしとありの対象者の平均予測確率の差と同等と解釈できる。

#### *Box H.*　Cox回帰モデルの性能評価

回帰モデルをもとにした大部分の予測モデルにとって、性能評価は単純である。例えば、ロジスティック回帰では、アウトカムイベントの実測の確率と予測リスクに応じた複数のグループ（通常10）の予測確率をプロットすることがよく用いられる方法である（**Box G**とitem 10d、15aを参照）。このような「較正プロット」で、モデルの識別能はリスクグループの予測確率の拡がりでも表現される(417)し、正式な識別能指標も計算することができる。（item 10d）。

時間ｰイベントデータのパラメトリックモデルにも同様のアプローチは適用することができるが、これらのモデルは滅多に使われない。そのようなデータには、ほとんどCox回帰が用いられるが、Coxモデルの較正の評価はそれほど単純ではない。というのは、Coxモデルは完全に規定されていないからである。このモデルは様々な特徴の患者間でリスクの相対的差を推定するが、ベースラインの生存関数を推定するわけではなく、絶対リスク（イベントの確率）を推定するわけではないからである(309)。このことの例外は、Coxベースの予測モデルの焦点がある特定の時点のアウトカム（例えば10年間の心血管死のリスク）にある場合である。この場合、関心のある時点のベースラインの生存確率だけが必要なので、識別能や較正は**Box G**に述べたようなやり方で評価することができる。

**COXモデルからの開発**

COXモデルは、複数の予測因子とそれぞれの回帰係数（対数ハザード比）の組み合わせで規定される（411, 531）。予測指標は、モデル中の変数の重みづけ線形和であり、その重みには回帰係数を用いている（item 15bの例参照）。個人の予測指標は、理論的に予測指標が0とした人との相対的な対数ハザードである（309）。

COX回帰を用いて新しいモデルを作成すると、予測指標（PI）はリスクグループ毎の予測生存時間を調べるのに利用できる。たとえば、患者をそれぞれのPI値をもとに4つの均等なグループに分類することができる。識別能はこれらのリスクグループごとのKaplan-Meier曲線を描き視覚的に確認できるし、数学的な評価も行う事ができる（item 10d参照）。較正は、COXモデルから直接得られた生存時間曲線を重ね合わせて考察する事ができる（309, 373）。

**COXモデルの検証**

実際、COXモデルでは基準生存関数は決して公表されない。その結果、絶対リスクは決して推定できないので、結果的に、他の研究者がCOXモデルの外的妥当性を検証することはできない。特に較正は、容易には評価できないだろう。RoystonとAltman（309）は、モデル作成の研究から入手可能な情報の量に応じた解析手法をいくつか提案した。

作成したモデル中の予測因子の正確なコードの方法、予測因子と対応する回帰係数が分かる場合に限り、識別能を調べる事ができる。そうすると、検証データセットのなかの個々の対象者の予測指標値を計算する事ができ、それを用いて、予測指標値を単一の共変量として回帰分析を行う事ができる。同じような診断関連群を集めて、検証データセットで識別能を測る事は、モデル作成データセットにおいて、予測指標に対する回帰係数をおよそ1とした場合とほとんど同じことである。もし検証データで傾きが1未満の場合は識別能が不良、逆にもし傾きが1より大きい場合、識別能は良好とされる。

さらに、もしモデル作成研究でKaplan-Meier曲線がいくつかのリスクグループ毎に描かれていれば、モデル作成データセットでの曲線とモデル検証データセットでの対応するKaplan-Meier曲線を比較すると、モデルの較正を大まかに評価できる。もし2セットの生存曲線が良く合致した場合、較正が良好と推察されるかもしれない（判断であり、正確な比較ではない）。このような較正の評価の仕方は、観察された値と予測される値の厳密な比較ではない。しかしながら、COXモデルは直接的に生存確率を予測することができないのである。基準生存関数がない状況では、個別のサンプルの較正がどれくらい良いかどうかは判断する事ができない（309）。

**2. 追加予測因子の増分価値を定量的に評価する**

*例*

我々は、バイオマーカーをGRACEスコアに追加するとどのくらい予測価値が増加するのかを尤度比検定を用いて評価した。3つの補助的な識別能改善を判断する計算を行い、それらは個々のバイオマーカーがGRACEに追加された際に、どれくらいモデルの性能が増加するかという大きさを求めた: 具体的にはAUCの増加量(△AUC)、IDI、連続およびカテゴリーでのNRIを計算した。臨床的な有用性を考慮し、我々はNRI（>0.02）を計算した。これは、予測リスクとして2%を臨床的に意味のある変化として、最低限の閾値としたということである。さらに、2つのカテゴリーNRIには、先行研究から事前に定めた6%と14%のリスク閾値を適用した。また、本研究の観察イベント発生率から5%と12%を適用した。カテゴリーNRIは、予測リスクがあるカテゴリーから別のカテゴリーに移動した際に、上方または下方に再分類された際の改善度と定義される。GRACEに追加されたバイオマーカーの数が少なかった（最大2個）ので、過度の楽観の程度は小さそうだった。しかし、我々は、△AUC とIDI解析をブートストラップを用いて内的検証し、結果を確認した（338）。[予後; 増分価値]

*解説*

単一のマーカーや検査の研究と比較して、多変量解析を用いる利点は、ある検査やマーカーが（他の要因を調整して）どれくらい増分価値があるか、直接的にエビデンスを得られることである。しかしながら、これまでに確立された複数の予測因子や現存する予測モデルに対して、ある新しい予測因子を追加する事で、どれくらい増加や改善があるかという従来から一般的な性能評価基準（識別能、較正、R2値など）を用いて、増分価値を定量化する事は困難であり、それを臨床的に解釈するのは難しい（339,340）。さらに、そのような性能評価のうちc統計量は増分価値を評価するのに感度が低いとの懸念がある（341, 342）。ただし、記述統計量としての役割は十分にあるとされている（343）。最後に、統計学的有意差検定は誤解を招く可能性がある。なぜなら、新しくて弱い関連の予測因子であっても、サンプルサイズが大きければ容易に統計学的に有意な相関を認めるからである。

したがって新しい測定方法は、事前に設定されたリスク分類に応じて、個々人をどのように再分類するかという考え方に基づいて提案されてきた。ここで再分類に用いる表とは、モデルにある予測因子を追加する場合とそうでない場合に、どのように個々人が再分類（低から高リスクへ、またその逆も然り）されるかが分かる表のことである（344-346）。ただし再分類の表は、リスクグループを定義する閾値に鋭敏に依存する（item 11）。

NRIは、そのような表で、再分類による増加分を定量化するのに現在よく使用されている測定方法である（339, 347, 348）。NRIは、モデル開発に使用する際には、ある予測因子を確立された複数の予測因子に追加したり、既存のモデルに組み込んだり（つまり、モデルが入れ子になることになる）する際に使用でき、モデル検証に使用する際には、入れ子になっていないモデル同士を比較する際に利用できる。それらは、比較したモデルが十分にデータに合うように調整された事を示すのに使用される（349）。そのため、NRIを使う前に、NRIの利用が適切であることを読者が判断できるようにモデルの較正を先にやる必要がある。

NRIはリスク分類を定義する閾値の影響を強く受けることが示されている（そして、それゆえ操作的である）また他にもいくつかの他の使用上の注意がある。特に、最適ではない較正を行ったモデルについては注意を要する（350-356）。したがって我々は、NRIを計算する際、必ず、関心のあるアウトカムの有無別に、対象者を層別した分類表と一緒に用いる事を推奨してきた（357）; item　 16。連続性NRIに関しても同様の懸念があり、モデルの改善よりむしろ関連を評価していて過大解釈となりうるとか、モデルの誤った較正につながるとされてきた（346）。

他の測定指標、例えば、純便益{net benefit}の変化や相対有用性{relative utility}の変化、重み付きNRIの改善度などがNRIより好ましいのではないかと提案されてきた。しかしこれらの三者は数学的に相互変換する事ができる（349）。あるひとつの予測因子を現存する予測モデルに追加する事で、増加する価値を定量化するのに最適の指標を定める事は、いまだに活発な研究領域であり、モデルに基づく尤度比検定による、臨床的で直観的な計測方法を見つける事は魅力的である（343, 358）。

システマティック・レビューによると、再分類に関する研究において、リスク閾値の選択に関する根拠となる文献はほとんど示されていないことが分かっている（105）。さらに、過半数の研究で、較正に関しては報告できておらず、再分類に関する適切や不適切の割合に関する情報はほとんど与えられていない。

**3.有用性の測定**

*例*

我々は、決断曲線分析（観察対象者の打ち切りを考慮した）を用いて、QRISK2-2011とNICEフラミンガム方程式の臨床的効果を記述し、比較した。個人が高リスクとされる閾値の幅に渡って純便益が最も高ければ、モデルの臨床的な価値があると考えられる。簡単に言うと、モデルの純便益は、高リスクに割り振られる閾値のオッズにより重み付けした真陽性の割合と偽陽性の割合の差である。一旦ある閾値が決まると、純便益が高いほどより良いモデルである（117）。[予後; 検証]

*解説*

識別能や較正は予測モデルの性能を表す統計学的な特性であるが、ある特定の段階の識別能の臨床的な転帰や誤った較正の程度を捉えている訳ではない（359, 360）。新しい方法、例えば決断曲線分析（361-363）や相対有用性（364-366）は、臨床的な帰結に対する洞察を与える。また、特定の閾値での予測モデルを用いた純便益に関しても同様である（349）。それらはまた、異なるモデルの臨床的な有効性を比較するのにも利用できる。具体的には、基本モデルと拡張モデルを同じデータセットに当てはめたり、2つの異なるモデル（別々の2つのデータセットから開発した）を同じ独立のデータセットで検証するのに利用できる（367）。

### Item 10e. 検証の過程で、もしなされているのであれば、モデルのアップデート（例: 再較正）について記述する。[V]

*例*

最初の候補として25個の診断予測因子があったが、36症例しかなかったので、[最初の診断モデルの] 係数は過剰適合になりやすい。そこで過剰適合の程度を定量化する為に[我々の検証用データセットで]ロジスティック回帰モデルを当てはめ、較正の傾きbを推定し、シュリンケージ因子を決定する。: Logit(P(Y=1))=a+b\*logit(p)　ここでは[Y=1は我々の検証用データセットにおける肺炎（アウトカム）ありを示す] pは予測確率のベクトルである。線形予測因子の傾きbはシュリンケージ因子とする。よく較正されたモデルではbは1である。このように、我々は最初のモデルの係数にシュリンケージ因子を乗法し係数を再度較正する（推定後の縮小）（368）。[診断; モデル更新、ロジスティック]

この研究では、我々はVan Houwelingenにより提案された”較正による検証”の[モデル更新の]方法を採用した。[原型の]UISS予測モデルを用いて予測した全生存値を、それぞれのリスクカテゴリーにおいてワイブル比例ハザードモデルを当てはめた。観察されたKaplan-Meier曲線にこれらの予測曲線を重ねて描き、それらの差を”較正モデル”で評価した。それは、3つの異なるパラメータ（α、β、γ）を調べる事で、原型の予測スコアがどれくらい新しいデータに対して妥当かを評価した。もしα=0、β=-1、γ=1という同時帰無仮説が棄却されるならば、（つまり、観察された曲線と予測した曲線に差があれば）較正モデルの推定が予測確率の再較正に使われる。再較正はモデルの識別能の精度を反映しないことに注意すること。この方法の詳細は Van Houwelingen とMiceliの文献に報告されている（369）。[予後; モデル更新; 生存]

外的検証の結果、モデルの更新を行った。我々は、予測モデルの切片と回帰係数をアイルランドの状況に調整した。オランダの環境と最も重要な違いは、移植に対するヘモグロビンの閾値の下限である。それはアウトカムと、以前のヘモグロビンレベルの予測因子に対する区分線形関数の区分点に反映される。二つの方法を更新に用いた。: モデルの再較正とモデル改訂である。再較正は、切片の調整と同じ因子を用いた個別の回帰係数の調整、つまり較正傾き、を行った。改訂されたモデルでは、個々の回帰係数は個別に調整された。これは一つずつ再較正モデルに予測因子を追加していく逐次追加法で行われ、変数が追加される度に尤度比検定（p<0.05）で評価された。予測因子に対する回帰係数は十分に調整された（370）。[診断; モデル更新; ロジスティック]

*解説*

既存の予測モデルを別のデータセットで検証（または適応）すると、モデルを開発したデータセットで推定された予測性能よりも一般的に落ちることになる。より厳密な検証方法で行えばその差はずっと大きくなり得る（Box CとFigure 1）。: 性能の低下は、同じ研究者による時期を変えた検証よりも、異なる土地や環境での検証、異なる研究者による検証の方がより起こりやすい（2, 20, 21, 102, 290）。予測精度の低下がみられたら、研究者たちは、単純にそのモデルを拒否するか、検証データセットに再度適合させるか、または新たに全く新しいモデルを作るだろう。

同じアウトカムや標的集団に対して全く新しいモデルを開発したくなっても、それは多くの理由で不適当なことである（20, 31, 102, 290）。まず初めに、時期ごと、病院ごと、国ごと、環境ごとに異なるモデルを開発する事は、予測研究を限定的にする。第二に、医療従事者がどのモデルを実際の現場で使うか決めにくくなるだろう。第三に、検証研究はしばしば対応する開発研究に比べ対象者が少ないので、新しいモデルは過剰適合しやすく、おそらく、原型のモデルに比べて一般化可能性が低くなる。最後に、原型の（モデル開発の）研究で得た知見が適切に使用されていない、それは、エビデンスに基づく医療が、出来るかぎり多くのデータによって得られた根拠によるという概念やガイドラインに直観的に反している（371）。

手元にある検証用のデータで新しいモデルを作成する前に、むしろ先に、どの程度の予測精度の欠落があり、それを補いうるのかを確認する為に、原型のモデルを調整（つまり、更新）する方が良いかもしれない（85）。調整したモデルは、原型のモデルに寄与した情報と検証用データの個人の情報を統合するので、他の対象者への転用可能性が改善するかもしれない。

予測モデルの更新の方法はいくつかある（2,20,31,102,290,372,373）。方法は広範囲にわたって異なる、そのことは再推定するパラメーターの数に影響される。一般的に開発と検証のデータセットにおけるアウトカム発生の割合は異なり、新しいデータでの原型のモデルの較正は良くない。検証用のサンプルに対して、原型モデルの切片や基準ハザード関数（もし分かれば）を調整すれば、較正は大抵よくなり、たったひとつのパラメーターの更新で済み、検証用データ数は少なくて済む（31,290,372,373）。より大規模なモデル更新の方法としては、単一の再較正因子の重みを用いた全係数の調整、特定の予測因子の重みの調整、新しい予測因子の追加から、すべての対象者個人の回帰係数の再推定まである。最後の方法の場合は、検証用のデータセットが開発用のデータセットより十分に大きい必要がある。

**Table 3**にはその他にさまざまなモデル更新の方法がまとめられている。単純なモデル更新の方法（1や2）は、単にモデルの較正を改善するだけである。識別能を改善する為には方法3から6が必要である。更新されたモデルでは、特に比較的小さな検証用データセットで行われた場合、実際の診療に適用する前にもう一度検証する必要がある（20）。

最後に、**Box C**に触れられているが、モデルの予測の性能を定量化する前に、新しいデータセットで現存するモデルを更新する事は推奨されていない（47）。もしモデルが更新されているのであれば、著者らはどのように行ったのかとそうした理由をきちんと述べるべきである。

## リスク群

### Item 11. リスク群が作成されていれば、どのように作成されたか詳細を記述する。[D; V]

**例**

最終モデルを定義した後、患者を2つの方法でリスク群に分けた。

1つは低、中、高リスク群の3群に分ける方法である（モデルのリスクスコアの分布の25パーセンタイルと75パーセンタイルにカットオフを設ける）。もう１つはCoxのカットポイントを用いて10群に分ける方法である。後者は与えられた数の群についての情報の喪失を最小限にする。臨床現場では3群に分ける方法の方になじみがあるので、ここからはモデルを特徴づけるのに3群に分ける枠組みを用いる（374）。[予後; 開発; 検証]

このモデルの目的の1つは臨床医が頭頸部癌手術を予定している患者のリスクを層別化するのに簡便に利用できる方法を開発する事である。このため、我々は輸血リスクを以下の3群に分けた: 低リスク群（15%以下）中程度リスク群（15%から24%）、高リスク群（25%以上）（375）。[予後; 検証]

NICEのガイドラインに沿って、10年間の心血管疾患の予測リスク20%以上の場合を高リスクとした（117）。[予後; 検証]

予後指標の分布の三分位に基づき、3つのリスク群に分けた。低リスク群（第１三分位、PI≦8.97）は5年及び10年無イベント生存率（event free survival: EFS）がそれぞれ100%及び89%（95%信頼区間、60−97%）だった。中程度リスク群（第2三分位、8.97＜PI≦10.06）では5年及び10年EFSは95%（95%信頼区間、60−97%）及び83%（95%信頼区間、85−98%）だった。高リスク群（第3三分位、10.06＜PI）では5年及び10年EFSは85%（95%信頼区間、72−92%）及び44%（95%信頼区間、24−63%）だった。[予後; 開発]

最後に、0%から100%の形で心不全の有無の確率を推測するために、縮小し、整数化した多変量解析の係数から診断尺度を開発した。心不全と判断したり除外したりするスコアの閾値は臨床的に許容できる診断偽陽性率（20%と30%）と偽陰性率（10%と20%）から算出した（377）。[診断; 開発; 検証]

**解説**

　多くの予測モデルの研究では、リスク群は多変量予測モデルの確率を用いて規定される。多くの場合これらは、結果の一部や臨床の意思決定の助けとして、例えば低、中、高リスク群のように名付けられている（item 3aと20）。

　どうやってリスク群を規定するのかという点や、実際いくつの群に分けるのかという点については一致した見解は無い（43）。もしリスク群を規定したら、用いた境界（つまりそれぞれのグループの予測確率の範囲）やどうやってその境界を選んだのかをはっきりと明記しなければならない。しかしながら、もし群分けが意思決定の助けになるようにという目的で規定されたなら、リスク群の数とその閾値の選択の意義について説明しなければならない。

　リスク群に分けることは患者にとって最大の関心事ではないかもしれないという意見もある（2、112）。このような群分けは、恣意的であり、論理的根拠が無いにも関わらず標準化されている場合もある（例えばNottingham Prognostic Index［378］が挙げられる）。また、予測の単純化はリスク（確率）がそれぞれの群の全ての個人において同等であるということを意味する。そのためリスク群の作成に関わりなく、群に基づくリスクだけでなく、個人のリスクも計算できるようにするために、論文中に十分な情報（ロジスティック回帰モデルの切片とベータ係数、ノモグラム、もしくは詳細な、より複雑な計算のためのWeb計算機）を記載すべきである(item 15a)。

　個別のリスク閾値（例えば、予後リスクがある閾値以上もしくは以下である場合に、心血管病変のアウトカムの予防のためスタチンが適応となるか否か[117]）に基づいた異なる治療や管理の計画を提唱する外部の知識に基づいてリスク群が規定されていることもある。しかし、多くの場合はそのような予測確率に基づいた明確なガイダンスは存在しない。

　癌領域における47の予測モデルのレビューによると、36（76%）の研究でリスク群を規定していた。しかし、そのうち17（47%）の研究では、どのようにリスク群を規定したのかは不明瞭か、もしくは報告されていなかった（54）。その他のレビューでも同様の事が指摘されている（43）。

開発対検証

### Item 12. 検証に際しては、開発とセッティング、組み入れ基準、アウトカム、予測因子の違いを明らかにする。[V]

**例**

・・・積算GRACEリスクスコアは退院から6ヶ月以内の全ての原因の死亡の予測確率に該当する。・・・その6ヶ月以降の妥当性は確立していない。この研究では、退院時のGRACEリスクスコアが退院後4年までの長期間における死亡を予測するかを別のコホートで検証する（379）。[予後; 異なるアウトカム]

Wellsスコアは深部静脈血栓症疑いで二次医療機関の外来に紹介されてきた患者から収集されたデータに基づいていた。二次医療機関の外来患者はプライマリ・ケアの患者と類似しているともしばしば言われているが、プライマリ・ケア医からの紹介メカニズムによる違いもあるだろう。Wellsスコアの真の診断もしくは識別能の正確さが、プライマリ・ケアにおける深部静脈血栓症が疑われる患者に関して正式に妥当性を検証されたことはない。どの診断もしくは予後予測ルールも新しい患者、特に他のセッティングから選択された患者に適応する際には元々の研究のデータから予測されるよりも予測能が低くなる傾向にあるので、妥当性検証の研究が必要である。我々はプライマリ・ケアにおけるWellsスコアの診断予測能を定量化し、元々のWellsらによる研究で報告された結果と比較することを試みた（188）。［診断; 異なるセッティング］

異なる研究間で、変数の定義が一致しない場合（例えば身体活動性など）、我々はデータベース間で納得できる一貫性を満たすために、利用可能な変数の内最適なものを採用するようにした。例えば、NHANESでは、我々は、「自身の活動性を、同年齢の人々と比べてください」という質問に対し「他の人より活動的」と答えた人を「活動的」と分類し、そうでなければ「非活動的」と分類した。ARICでは、身体活動性は「はい」、「いいえ」で回答する質問で評価したが、CHSでは身体活動性を、「全くない」、「低い」と「中程度」、「高い」の２つに分類した（380）。［予後; 異なる予測因子］

NWAHSでは降圧薬の使用に関するデータを収集していなかったので、降圧薬を使用した対象者はいなかったものと仮定した。同様に、BMESでは高血糖の既往に関するデータを収集していなかったため、誰もそういった既往がないと仮定した（381）。［予後; 異なる予測因子］

**解説**

予測モデルの異なるデータセットにおける予測能を評価する研究においては、著者は外挿可能性に影響を与え得るどのような差異も明確に記載しなければならない（26, 28）。

ある一つのセッティング（プライマリ・ケアのような）やある特定の国で開発された予測モデルは他のセッティング(二次ケアのような)や他の国においては必ずしも同等に利用することはできない(19–21, 26, 28, 33, 183, 382, 383）。例えば、診断関連群（item 5a）はプライマリ・ケアと二次ケアで異なる傾向にある。両者を比較すると、二次ケアではより多くの徴候や症状（そして予測因子の取る値の幅は狭い）とより進行した病状が認められる(20, 21, 102)。

組み入れ基準は意図せず異なり（例えば、より広いもしくは限定された年齢層など）、診断関連群において差異を生み出すこともあるし（186）、意図して異なる場合もある（例えば、成人で開発された予測モデルを小児で検証を行う場合など［191, 384］）。

検証研究のアウトカムは、開発研究のアウトカムと同一に見えるかもしれないが、正確な定義や測定方法は異なることがあり得る。例えば、糖尿病は随時血糖、経口糖負荷テスト、自己申告のいずれを用いても定義できる（380, 385）。アウトカムの定義と測定が同一であったとしても、状況の違いにより差異は存在するかもしれない。例えば、観察者の専門度（放射線科医や病理医といった）、異なる検査法や異なる画像技術が原因となり得る。

セッティングや適格基準と同様に、アウトカムの差異も意図的なものである場合がある。研究の目的が異なるアウトカムを予測するのに利用できるかを評価することであることもある(379, 383, 386)。心臓手術の術後死亡率を予測するために開発されたモデルは、集中治療室在室期間を予測できるかについても評価された（46）。既存の予測モデルが別の時期の同じ結果を予測できるかについても評価されている。（387）。例えば、急性冠症候群患者の6ヶ月死亡率を予測するGRACEモデルは（388）、4年時点での死亡率を予測できるかについても評価された。（379）。

最後に、予測因子の定義と測定法は、繰り返しになるが意図的であろうとなかろうと、異なる場合がある。定義が同じであれば予測因子が測定された状況の違いにより差異が生まれるかもしれない。例えば、血液中のある成分はもともと検査室で静脈血を用いて測定されるものかもしれないが、ベッドサイドの毛細血を用いた簡易的検査でも妥当性を検証されている（136, 389）。

検証研究の著者は、予測因子がどのように変数化されたのかも明確に報告しなければならない。これには、全ての連続変数の予測因子の単位を記したり、どうやってそれぞれのカテゴリカル変数が変数化されたのか（例えば、性別で言えば、女性は0、男性は1といったように）報告したりすることも含む: item 7aと15aを参照。さらに、予測モデルの予測能を評価するのに過去のデータを用いている場合、データがそもそも別の目的で収集されていて、予測因子のデータが取られていない場合もある。研究者は近似的な予測因子を用いたり、代入したり、除外することになる（198）。予測因子を除外する（つまり0として値を代入する）ことは、検証ではモデルの予測能の解釈が困難になるので避けなければならない（198）−例としてFRAXがある（312, 314）。

そのため、検証研究の著者が、セッティングや適格基準、アウトカムの定義や測定法に（意図的であろうとなかろうと）変更があるかどうかを記したり、状況、定義、測定方法が開発研究と同等である旨を記載したりすることは重要である。著者は適格基準、アウトカム、予測因子を挙げるだけでなく、差異やどうそれらを取り扱ったかについても明確にしなければならない。

　直近の外的検証研究のシステマティック・レビューでは、45研究の内6研究（13%）で結果の定義が元々の研究の定義と同等であるか明らかではなかった（122）。

## 結果

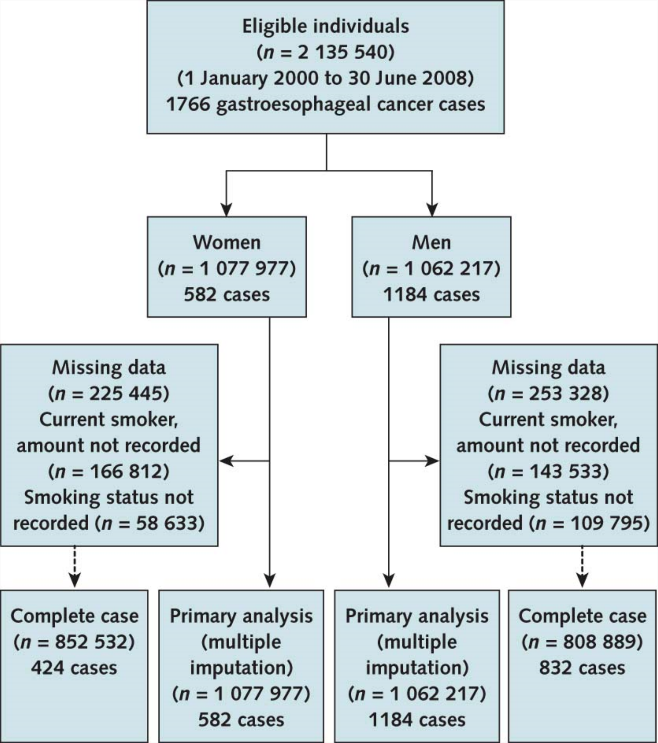
## 対象者

### Item 13a. 対象者のアウトカムの有無と人数、可能であればフォローアップのまとめを含めて研究を通じた対象者の流れを記述する。ダイアグラムが手助けになるだろう。［D; V］

例:対象者の流れ

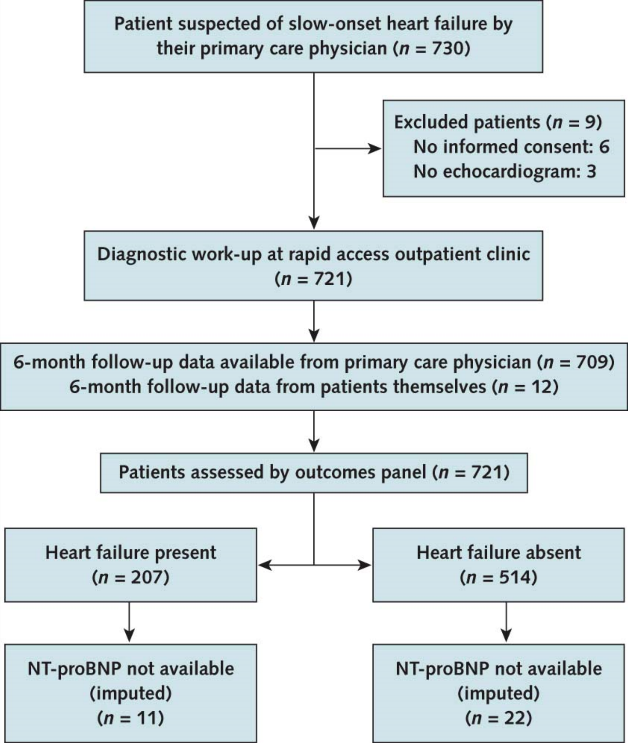
**Figure 3, 4**参照

#### Figure 3. 図の例：対象者のフロー図



引用文献 390よりElsevierの許諾を得て転載

#### Figure 4. 図の例：対象者のフロー図



引用文献 377より許諾を得て転載。NT-proBNP = N-terminal pro-brain natriuretic peptide.

**例: フォローアップ期間**

我々はTHINコホートの全ての患者の10年間の予測心血管リスクをQRISK2-2011リスクスコアを用いて計算した。そして292,928人（14.1%）を10年以上追跡した。［予後; 検証］

解析時点で、204人（66%）の患者は死亡していた。生存している患者のフォローアップ期間の中央値は12ヶ月（範囲1-84ヶ月）だった（391）。［予後; 開発］

フォローアップ期間の中央値は“reverse Kaplan Meier法”により計算した。これは通常のKaplan Meier法と同様の方法で潜在的フォローアップ期間を計算するが、打ち切り・イベントの指標を逆にしている。従って、死によって個人の真ではあるが知ることのできない観察期間は打ち切られ、打ち切りがイベントとして扱われる（Schemper & Smith, 1996）（392）。［予後; 開発］

**解説**

　どのように対象集団が選択されたかも含め、研究対象の母集団について読者が理解できることは重要である。こうした情報は予測モデルが検証できるか、もしくは適応できるかを判断するのに不可欠である。研究における対象者の流れは文章もしくは表でも示すことができるが、フローダイアグラムは、モデルを開発や検証の研究サンプルの抽出法を明らかにするのに価値ある方法である。

　フローダイアグラムの始まりのポイントは対象者の元であり、その後に続く段階は適格基準やデータの利用可能性と関連する場合がある（108）（item 5b）。欠測データのある対象者の人数や、結果の発生数などその他の情報をダイアグラムに含むのも役立つ場合がある。

　予後予測の研究では、対象者のフォローアップ期間を要約することが重要である。これはしばしば時間の中央値で与えられる。フォローアップ期間の中央値の算出方法も規定すべきである。一つの方法としてreverse Kaplan Meier法があり、これはコホート内のすべての患者のデータを用いる（393）。ここでは、打ち切りをアウトカムとするため、イベント指標を逆にして標準Kaplan Meier法を用いる（108）。イベント発生しなかった患者（言い換えると打ち切りのある生存期間を持つ患者）のフォローアップ期間の中央値を記すことも役立つかもしれない。特定の時点でのイベントの確率を予測するモデルでは、その時点まで観察された対象者の人数を記載することも役立つ。

　遅れての疾病診断をアウトカムとする診断研究では（item 4aと6b）、フォローアップ期間の中央値は重要である。もし研究データを開発用と検証用に分けているのなら、それぞれにおいて前述の情報を記述したほうが役立つ。

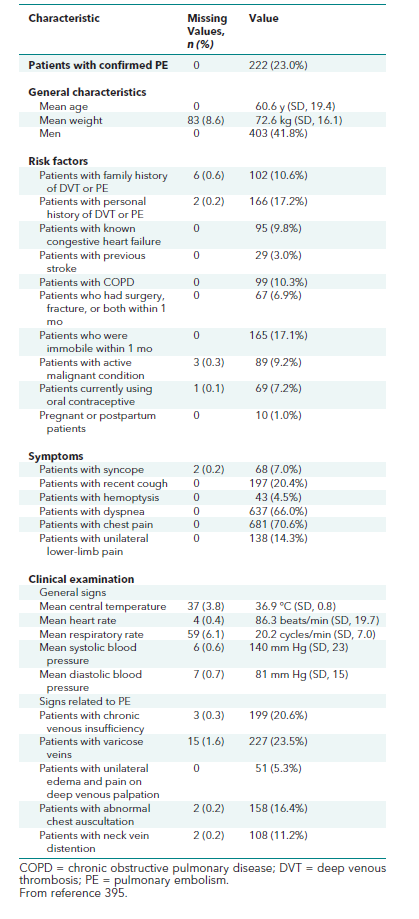
　最近の予測モデルのシステマティック・レビューでは多くがアウトカムのイベント発生数を明記していないことが分かっている（34, 45, 54, 85, 394）。他のレビューでは、フォローアップ期間の要約を記していない研究もしばしば認められることが分かっている（43）。

### Item 13b.予測因子とアウトカムについて欠測がある対象者を含めて、対象者の特徴（基本的な人口統計、臨床的な特徴、可能な予測因子）を記述する。[D; V]

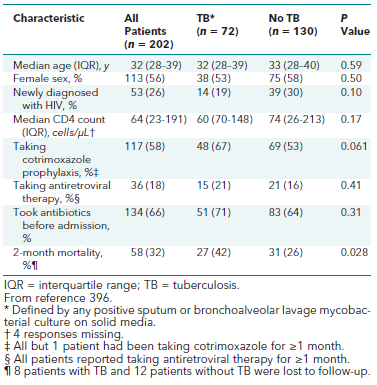
**例**

**Table 5, 6**参照

#### Table 5. 表の例：患者特性



#### Table 6. 表の例：患者特性



**解説**

　研究対象者の背景因子の分布（罹患率、平均または中央値と標準偏差、または四分位）の明確な記述は背景、診断関連群、セッティングを判断するのに重要である。読者は予測モデルが彼らのデータで潜在的に妥当かどうか、さらに彼らの患者にも適応できるかを判断できる。研究の組み入れ基準だけを報告するのでは充分でない。情報は、理想的には、全ての、特に最終モデルに含まれている、予測因子、及び他の重要な変数（人口統計属性、臨床、セッティング）について提供されるべきである。さらに全ての連続変数の予測因子の、特に最終モデルに含まれている因子の範囲についてはっきりと報告するべきである。予測因子の範囲が分かなければ、どの患者にモデルを適用したらよいのか分からない（item 15a）。

　上記の情報は表を使うと最も効率的に表現できる。これにはそれぞれの変数の欠測値の数（パーセンテージ）も組み込むべきである（Table 4）。もし欠測値が研究の変数のごく僅かにしかなければ、この情報は文章で要約してもよい。

　もし単変量解析を行ったのならば、異なるアウトカムカテゴリーことに、予測因子や他の重要な変数の要約統計量ついて記述することも役立つ（item 14b）。その代わりに、予測因子のカテゴリ別にアウトカムの頻度を示しても良い。

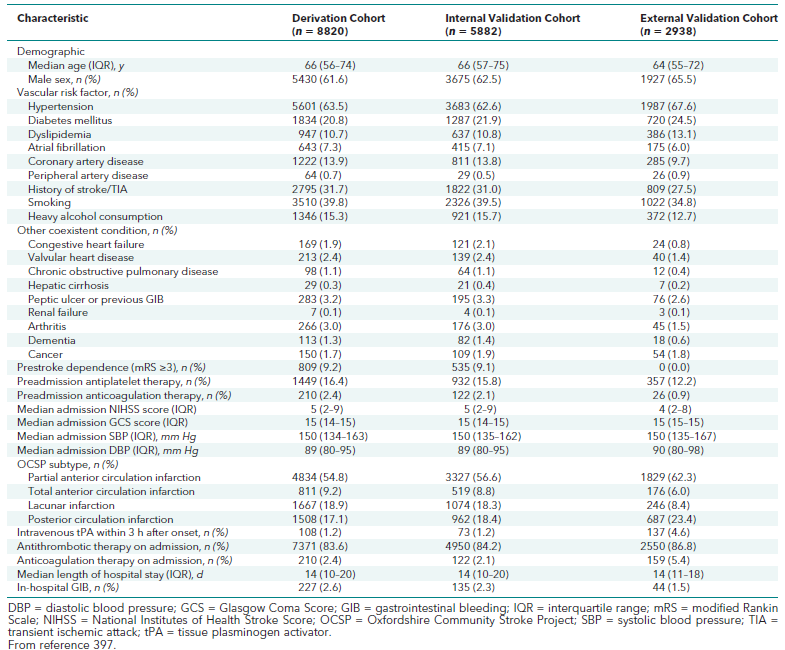
　対象者の特徴や予測因子を報告するのが特に不十分であるというエビデンスはない。しかし、いくつかのシステマティック・レビューはこうした重要な情報を報告していない研究の存在を明らかにしている(43, 62, 71, 72, 122)。120個の予測モデルの外部妥当性を評価した78個の研究の最新のレビューでは、連続変数の予測因子の範囲が示されていたのはもともとの予測モデルを開発する研究の内の8%（120の内の10）に過ぎなかった（122）。

### Item 13c　検証に際しては、重要な変数（人口統計、予測因子、アウトカム）の分布を含めて開発のデータとの比較を提示する。[V]

例

**Table 7**と**8**を参照

#### Table 7. 表の例：開発と検証データの患者特性の比較[開発; 検証]



#### Table 8. 表の例：開発と検証データの患者特性の比較[検証]

**解説**

予測モデルの検証研究は、一般的にモデルを開発した元の研究の対象者と類似した対象者で行う(19, 20, 26, 28, 33)。しかし、item 12で述べたように、検証研究では開発研究の対象者とあえて異なる集団を対象にすることがある。検証研究の対象者の人口統計属性、モデルの予測因子とアウトカムを元の開発研究で示されている対象者とともに報告することが重要である。そのような情報は、全標本および適切な要因（性別など）で分類された集団でそれらの変数の分布を表にすることで効果的に示すことができる。また、両方のデータセットでそれぞれの変数の欠測値の数も報告するのが有用である。

以前からよく使われているモデル（APACHEリスクスコアやフラミンガムリスクスコアなど）ではこのような比較は不要だという意見もある。しかし、必ずしも全ての読者がそのモデルや元の開発研究のデータと検証研究のデータの比較、もしくは先行する検証研究について知っているわけではないので、やはりこのような比較をすることを推奨する。

最後に、検証研究と先行研究の対象者に意図しない特筆すべき差があるときはその理由を説明すべきである（item 12）。そして、その検証研究でのモデルの予測精度などの観察された結果について適切な解釈を付け加えるべきである（item 16と18）。

最近のシステマティック・レビューによると、78編の外的検証研究（外的検証をしているモデル開発研究も含む）のうち31 編（40%）しか元のモデル開発研究と外的検証研究の対象者特性を比較したり議論したりしていない(122)。

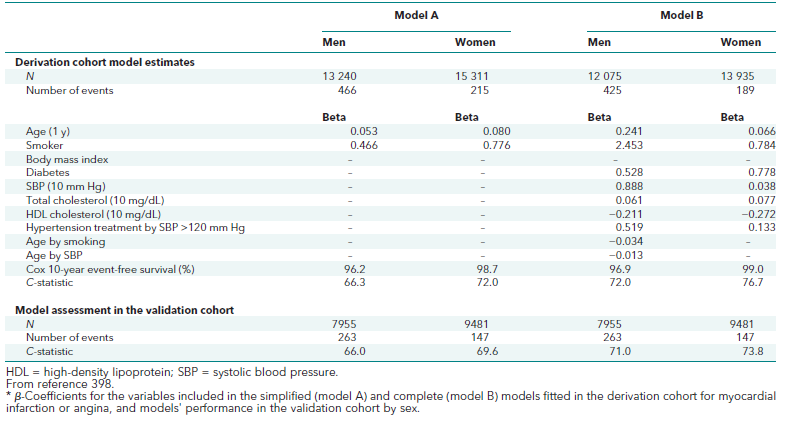
## モデルの開発

### Item 14a. それぞれの解析における対象者とアウトカムの数を明示する。[D]

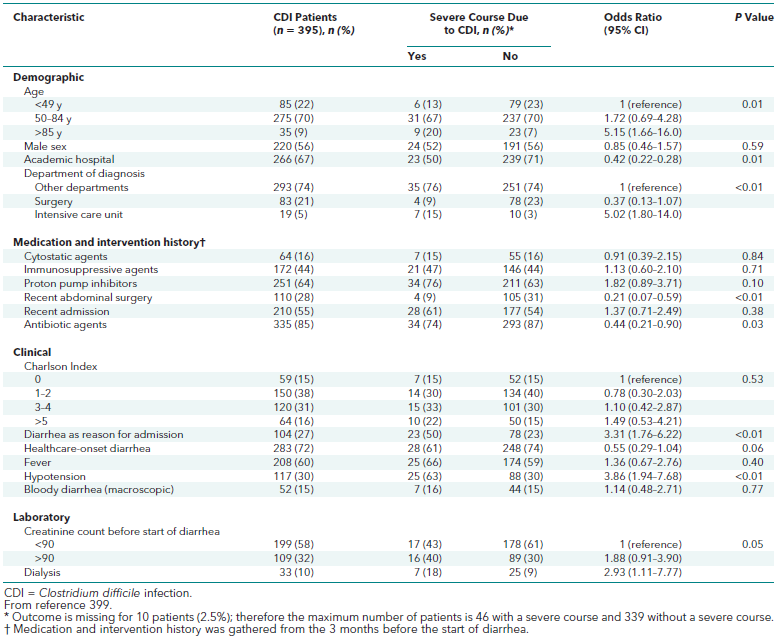
**例**

**Table 9**, **10**を参照

#### Table 9. 表の例：多変量モデルのサンプルサイズとイベント数を報告\*



#### Table 10. 表の例：それぞれの単変量解析でのイベント数を報告



**解説**

Item 8でも解説した通り、予測モデル開発研究で必要な標本数はイベントの数であり、対象者数ではない。検討された予測因子の数と関連したイベントが起きた対象者数は、その研究で過剰適合のリスクを評価する際に重要である（item 8と10b）。

欠測値があるデータの場合、ひとつでも欠測値がある対象者を除外するか欠測値の補完をしない限り、対象者数とイベント数は解析によって異なることがある（item 9）。新しい予測モデルを作るとき、1つの予測因子とアウトカムの未調整の関連（しばしば単変量または二変量関連と呼ばれる）を解析することが多い（item 14b）。このような場合、どちらか1つでも欠測値がある対象者を解析から除外する（ペアワイズ除去）ので、ひとつひとつの予測因子とアウトカムの調整しない関連で対象者数が異なる。そのため、単変量関連を報告する場合は、それぞれの予測因子で欠測値のない対象者数とその中でのイベント数を報告するべきである。

同様に、同じデータセットから複数の多変量モデルを作成して性能の比較をすることがある。あるモデルは日常的に測定している予測因子を用いて作り、別のモデルは血液検査結果のような日常的に測定していない予測因子も含めることがある。重要なのは、全てのモデルでそれを作るために組み入れた対象者数とイベント数をがわかるようにすることである。

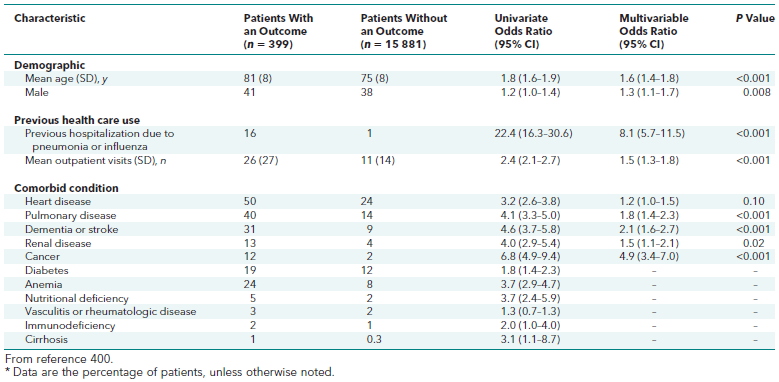
それぞれの解析にどの対象者が組み込まれたか、理解しやすく報告しなくてはならない。特に新しい予測モデルを作る場合、モデルを作成するために用いられたイベント数（EPVなど）を示すことで、過剰適合の指標を計算することができる（item 8と10b）。予測モデルを開発する研究で、データを開発と検証用のデータセットに分ける場合、それぞれのデータセットで対象者数とアウトカム発生数を報告することが重要である。

### Item 14b. もしなされているのであれば、それぞれの予測因子候補とアウトカムの調整抜きの相関を報告する。[D]

例

**Table 11**を参照

#### Table 11. 表の例：それぞれの予測因子とアウトカムについての未調整の相関\*



**解説**

読者が先行研究から予想されている予測因子を確認するためや、未調整（単変量）と調整（多変量）した解析の予測因子の予測精度の差を観察するために単変量解析を報告することが望ましい。これは、病因（原因）研究や非無作為化介入試験でいわゆる未調整や調整した相関を一般的に報告するのと同じ理由である(97, 401)。調整しない結果は、最終的な多変量予測モデルで調整した結果と比較するための基準となる。

二値アウトカム（例えば30日死亡）の単変量解析では、リスク比またはオッズ比を信頼区間とともに報告すべきである。同様に、イベントまでの時間をアウトカムとする場合、ハザード比と信頼区間を報告するべきである。P値も示してよいが、信頼区間以上の情報を加えるものではない。典型的には、このような解析結果はできる限り多変量解析による予測因子とアウトカムの相関の結果と合わせて表の形で報告するべきである。

欠測値がある場合、それぞれの未調整解析で組み入れた対象者数も示す（item 14a）。予測因子が二値またはカテゴリー変数の場合、それぞれの区分でアウトカムを発生した対象者が何人いたかも報告するべきである。

しかしながら、我々はそれぞれの予測因子とアウトカムの未調整の相関を元にして、どの因子を多変量モデルに組み込むかを選択するべきではないという、先行文献の著者の意見に賛同する(2, 112, 235)（item 10b）。

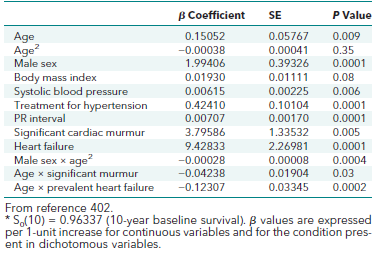
### モデルの仕様

### Item 15a.　個人での予測を可能とする予測モデルの全てを提示する（全ての回帰係数と切片や、任意の時点での基準となる生存率など）。[D]

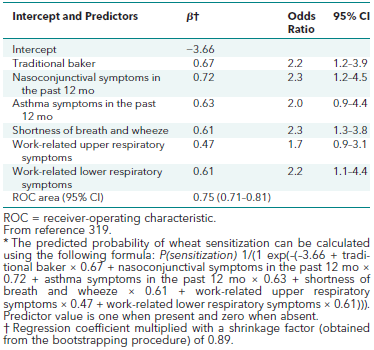
**例**

**Table 12, 13, 14**を参照

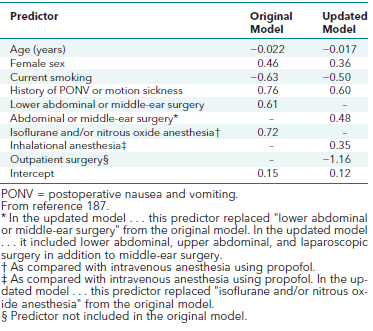
#### Table 12. 表の例：ベースラインでの生存確率を含んだ特定の時点\*での予後（生存）モデルを全て表示



#### Table 13. 表の例：切片\*を含んだ診断（ロジスティック）モデルを全て表示



#### Table 14. 表の例：元と更新された予測モデルの両方を表示



**解説**

予測モデルを検証研究または臨床での実践で個人の予測ができるように、十分詳しく報告すべきである（item 15b）。二値アウトカムであれば、モデルに組み入れたそれぞれの予測因子の回帰係数またはオッズ比と切片を示すということである。それぞれの係数の信頼区間を示すのが望ましいが(403)、それが検証研究や実臨床で使われることはない。イベントまでの時間がアウトカムの（長期）予後を予測するパラメトリック生存モデルでも同じように考慮する。シュリンケージ法を用いたときは（item 10b）、元の係数とシュリンケージ回帰係数を両方報告すべきである。

イベントまでの時間をアウトカムとしたよく見られるセミパラメトリックモデルであるCox回帰を用いる場合は状況がやや異なる。モデルに組み入れたそれぞれの予測因子の回帰係数またはハザード比を信頼区間とともに示すべきである。しかしながら、Coxモデルには切片はなく、個人の生存確率は不特定の基準点での生存時間関数を基に算出される。それゆえ、生存確率は回帰係数のみからは求められない。

ある時点の個人のアウトカム発生の確率を求められるようにするために、少なくとも1つの臨床上重要な時点での基準となる累積ハザード（または生存率）を報告すべきである（item 15b）。循環器や腫瘍の研究では、よく5から10年生存率を選択するが、その他の時点が重要であることもある。あるいは、Cox回帰モデルを用いて予測モデルを作る場合は、区分多項式や制限3次スプラインを用いて基準となる生存時間関数を求めて報告することを検討すべきである(297, 309, 373, 404)。

複雑なモデル（例えばICNARCモデル[405]）の詳細を全て提示することは容易ではない。他には、モデルを論文に発表するだけではなく、定期的継続的に更新されてウェブ上で公開されている場合もある（例、QRISK2 [139]）。モデルの複雑性や更新頻度に関わらず、査読された論文またはウェブ上の付録でモデルの全容を詳述することを強く推奨する。モデルの詳細が発表されないままだと妥当性の検証をすることができず、臨床で使えるかどうかが全く疑わしいものとなってしまう(312, 314, 406)。

開発したモデルの式を明確に報告することに加え、全ての予後因子がどのように数値化されたかをわかるようにすることが求められる（item 7a）。全ての連続変数の予測因子はその測定尺度を報告すべきである（例えば腰囲がセンチメートルかインチのどちらで測定されたか）。連続変数をカテゴリー変数に変換した場合（item 10aと**Box E**）、全てのカテゴリーでそのカットポイントを示すべきである。初めのカテゴリーの下限値と最後のカテゴリーの上限値が忘れられがちなので気をつける。カテゴリー変数の予測因子はどのように数値化したかを明確に示す（例えば、性別では女性を0、男性を1）。

さらに、全ての連続変数の予測因子の範囲も明確に示す。予測因子の範囲がわからないと、モデルを誰に適応できるかが定かではない。例えば、30から60歳の対象者を基にに開発された予測モデルを65歳の個人に当てはめても外挿していることになってしまう(186)。

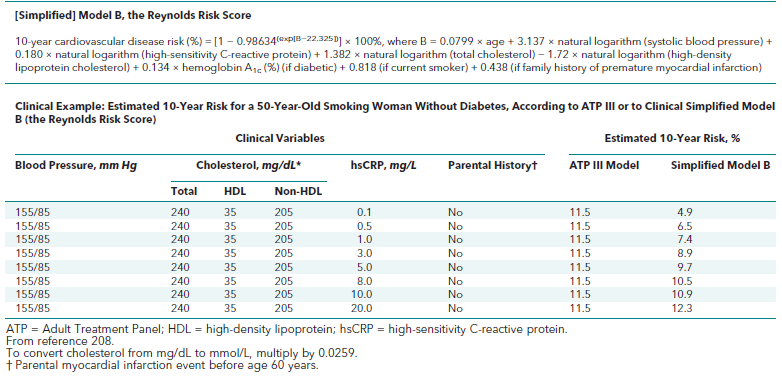
多くのシステマティック・レビューで、予測モデルを開発する研究では検証の研究や他の個人にモデルを当てはめるための情報が不十分であることが多いと報告されている(43, 62, 66, 88)。例えば、乳癌の予後予測モデルを開発した54編の研究のうち13編（24%）(43)、超未熟児の死亡率の予測モデルでは41編のうち22編（54%）(66)しか十分な情報を提供していなかった。他のレビューによると、レビューされた研究のうちでモデルに組み入れられた連続変数の予測因子の範囲を示していたものは1編もなく(53)、最近の2編のシステマティック・レビューでは年齢の範囲さえもしばしば示されていないと報告している(73, 74)。

### Item 15b. 予測モデルをどのように使うか説明する。[D]

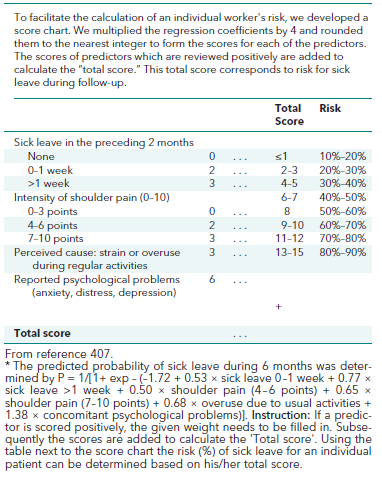
**例**

**Table 15から17**、**Figure 5から7**を参照

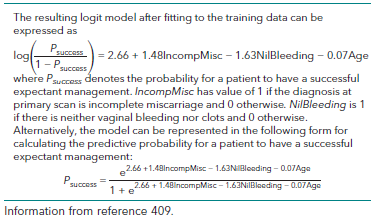
#### Table 15. 表の例：特定の時点でのベースラインとなる生存確率と、モデルを個々人での予測にどう活かすか示すための仮想個人でのデータを含んだ全てのモデルを表示



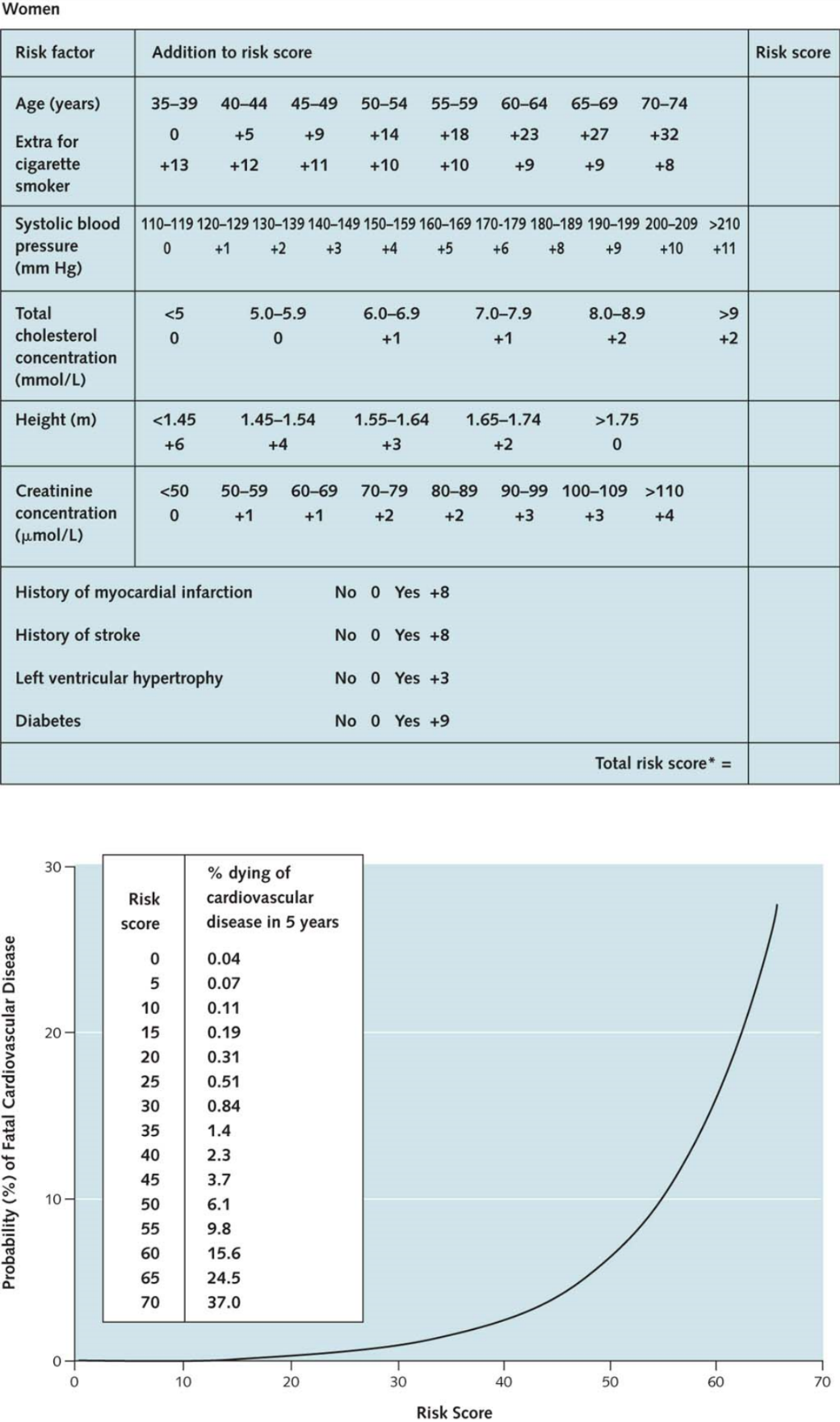
#### Table 16. 表の例：個人でのアウトカムリスク（確率）が得られる単純なスコアリングシステム\*



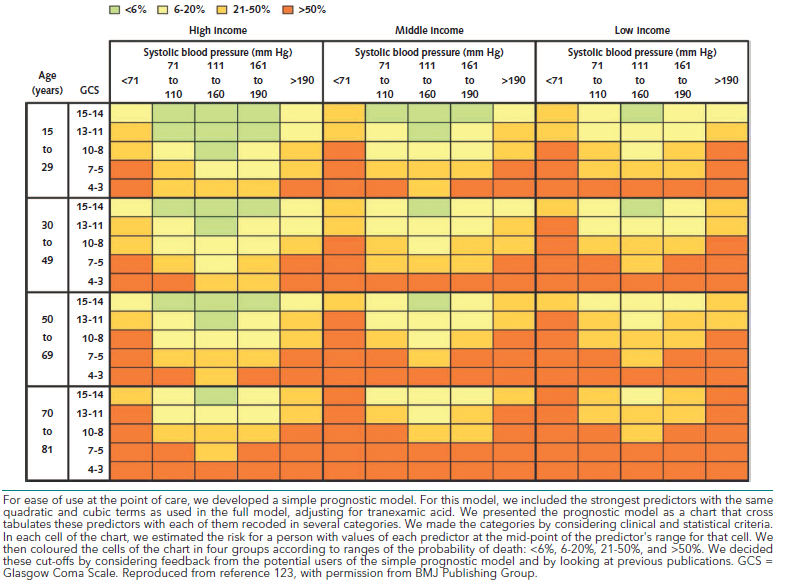
#### Table 17. 表の例：個人での予測確率を計算するための全ての詳細を提供



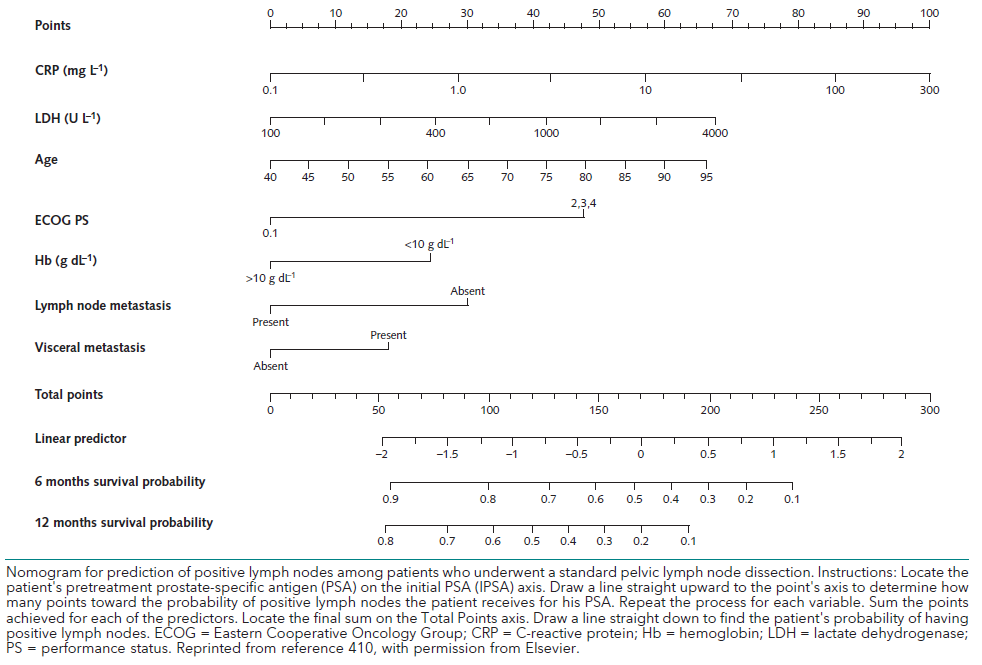
#### Figure 5. 図の例：個人でスコアごとに予測される確率を得るための図と合わせたスコアリングシステム

 **Reproduced from reference 408, with permission from BMJ Publishing Group.**

#### Figure 6. 図の例：個人での予想確率を得るための図を用いたスコアリングシステム



#### Figure 7. 図の例：ノモグラムと個人での予測確率を得る方法



**解説**

個人のアウトカム発生の確率やリスクを推定できるようにするために、開発した予測モデルをどのように使えばいいかを説明すべきである。回帰モデルでは、モデルの予測因子（測定されたものか数値化されたもの）の回帰係数を重みとした加重合計で線形予測が示される。（item 15a）。予後予測では、線形予測はしばしば予後指標と呼ばれる。ロジスティック回帰モデルの回帰係数はオッズ比の対数であり、Coxモデルではハザード比の対数となる。イベントまでの時間がアウトカムのCoxモデルを除いて、回帰モデルは切片(定数項)をもつ。

アウトカム発生の確率の予測は、予測因子の値のどの組み合わせからも求められる。ロジスティック回帰モデルであれば、アウトカム発生確率の予測は、βjを予測因子Xjの回帰係数、β0をモデルの切片として以下の式で表せられる。

probability = exp(β0 + β1X1 + β2X2 + ... + βkXk) / (1 + exp(β0 + β1X1 + β2X2 + ... +βkXk))

=1 / (1 + exp(−(β0 + β1X1 +β2X2 + ... +βkXk)))

これにより、読者にとってこの式が明確になる。確率に100をかけることにより、パーセントリスクとなる（%リスク = 100 × 確率）。

Cox回帰モデルによる予後予測モデルでは、ある追跡期間tでのアウトカム発生確率の予測は予後指標と「基準点での生存率」の推定S0(t)から求められる(112, 274, 411)。生存確率の予測はS0(t)exp(β1X1+β2X2+...+βkXk)で求められ、アウトカム発生確率は1- S0(t)exp(β1X1+β2X2+...+βkXk)となる。これらの値は100をかけることでパーセントに変換できる。

新しい予測モデルを作るときは、ベッドサイドモデルと言われるような、使いやすいように単純化したり得点法にしたりするものが多い(45, 53)。ここで言う「単純化や得点法」とは、item 10bで扱ったような予測因子を少なくしたモデルではなく、回帰モデルを基に単純に表現された式のことである。致死的心疾患の10年リスクを予測するSCOREモデル(140)や肺塞栓症診断のWellsモデル(412)のように、多くの有名なモデルが簡単に実践できるように変換されたり単純化されたりしている。例に示すように、アウトカムや生存確率を求めるために、最終モデルに組み入れたそれぞれの予測因子の回帰係数を計算しやすいように整数に変換（例えば四捨五入[413]）するなど、モデルを単純化する方法はいくつかある(414)。回帰係数を切りよくする極端な方法は、最終モデルに組み入れた予測因子全てに同じ重みづけをして、単純に持っているリスク因子の数を数えることである。簡単に計算できるスコアとそれによって求められるアウトカムの確率は、例に示すように表や図に表すことができる。

開発した予測モデルを四捨五入などで単純化することによって、予測精度はいくらか損なわれてしまう(1, 413)。よって、元のモデルの公式を単純な得点法に変換する場合、変換前後の予測精度の尺度（例えばcインデックス）（item 10dと16）を示すことが望ましい。それにより、読者が単純化されたモデルを使うことによってどの程度の予測精度が損なわれるかを判断することができる。もし単純化するのであれば、単純化した得点は元の回帰係数の基準（オッズやハザードの対数）を基にしなくてはならず、オッズ比やハザード比などに変換した係数を用いてはならない(415)。特に、オッズ比やハザード比が1以下となる予測因子（つまり、アウトカムに対して関連がないか負の関連をもつ）の場合、どのように得点を割り振るかを慎重に検討しなくてはならない。そういった場合に、正の得点を割り振ると合計点が増えてアウトカム発生の可能性が高く見積もられるが、実際は可能性を下げるように寄与するはずである。

単純な得点法のモデルを開発した場合、単純なモデルを作成した詳しい手順とそのモデルの得点がアウトカムの確率にどう関連しているかを明確に説明すべきである。得点法は、得点の取りうる範囲とそれに対応するアウトカムの確率を表や図で明確に示すことができる。単純化したモデルから得られた値を区分することで、リスクや確率の分類を作ることもできる（item 11）。この場合、そのモデルである範囲の値を持つ対象者は全て同じリスクの分類に入り、同じリスクを持つと評価される。しかし、単純に対象者のリスクを低い、中等度、高いと分類しても、それぞれの群の実際に予測されるリスクの大きさを評価しなければ不十分である。得点の区分は、観察または予測された、もしくはその両方のアウトカムの確率（の平均や範囲）と対応していなくてはならない。

生存モデルの場合、それぞれのリスク群でカプラン・マイヤー曲線を示すべきであり、それによって予後の差の自然な記述をすることができる（例、識別能）。カプラン・マイヤー曲線はそれぞれの群で患者の総数、アウトカムを起こした患者の数とイベント発生までの時間の概要（とその信頼区間）を示すことが望ましい。

予測モデルはしばしば計算図表（ノモグラム）として示される(310)。この表現方法は開発したモデルを単純化するものではく、元の数学的回帰式を視覚的に表現する方法である(112)。こういった形式は多くの読者や使用者に馴染みがないことがあるので、どのように計算図表を用いて個人の予測に適用するか、明確な説明をつけることが求められる。計算図表は回帰式を完全に説明できる方法ではない（item 15a）。

最後に、モデルをどのような方法で表すかに関わらず、特定の予測因子の状態を持つ仮想の個人を見立てた臨床シナリオを用いて、予測モデルを応用する例を示すのは有用である。

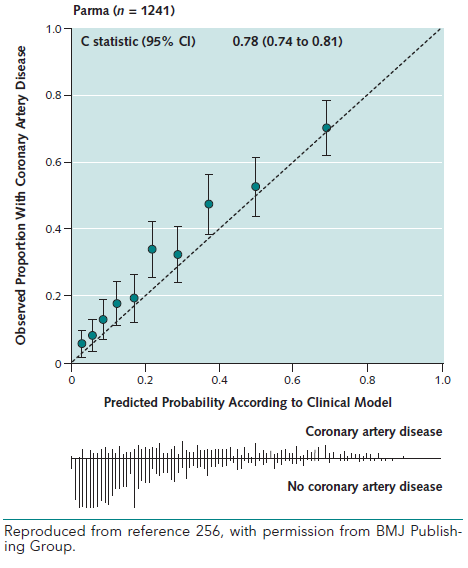
### モデルの性能

#### Item 16. 予測モデルの性能を示す指標を信頼区間とともに報告する。[D; V]

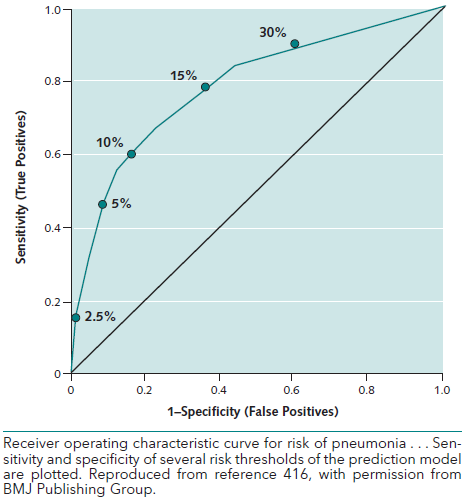
**例**

**Figure 8 - 10**と**Table 18**を参照

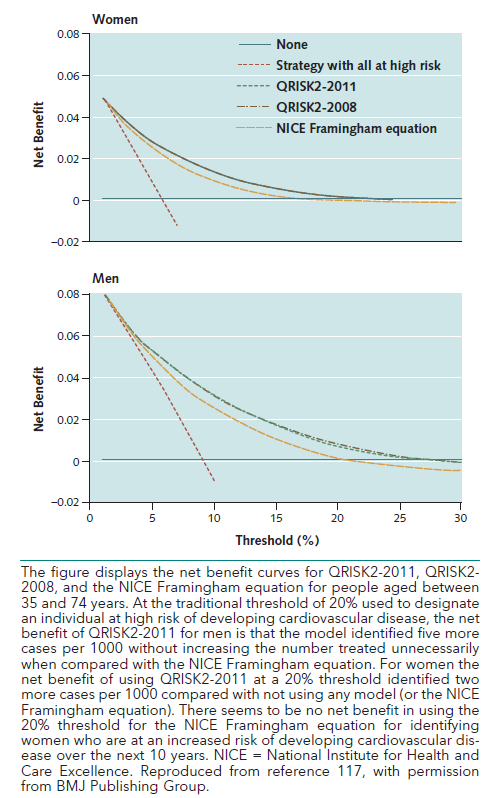
#### Figure 8. 図の例：較正プロットとc統計量、アウトカム(冠動脈疾患)のあり、なしに分けた対象者の予測確率の分布



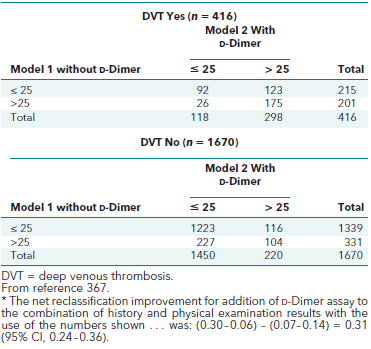
#### Figure 9. ROC曲線と予測確率を曲線に表示



#### Figure 10. 図の例：決断曲線分析



#### Table 18. 単一の閾値\*を用いた基本と拡張診断モデルについての純再分類改善度表の例（NRIと95%信頼区間）



**解説**

方法の項で記述されたすべての性能指標(item 10d)は、できれば信頼区間と共に結果の項で報告すべきである。複数のモデルの開発または評価をする場合には、それぞれのモデルの性能指標を報告すべきである. モデル開発研究において、内的検証の結果は、楽観を是正した性能指標(例: 見掛け上のcインデックスと修正されたcインデックス両者の報告)を含めて報告すべきである. Item 10bとBox F. 予測モデルが単純化されているなら(Item 15b)、単純化されたモデル同様にオリジナルのモデル(例: 全部の回帰モデル)の性能(例: cインデックス)を報告すべきである。

単一の数値化による報告に加えて、それぞれのグループごとにヒストグラムや密度プロット{density plot}、点プロット{dot plot}で個人ごとのアウトカムの有無を示して識別能をビジュアル化することが推奨される(417)。ロジスティック回帰モデルにおいてROC曲線があるが、曲線に適切なリスク予想を明確に示せていない限りにおいては、ほぼ情報量がなくc統計量以上の情報は示せない。

同じデータセットで2つ以上のモデルを開発するか(例: 基本モデルと拡張モデル[418])、評価する際には、そのモデルが同じデータで開発と検証されることを考慮に入れた統計手法が必要となるが、それらのモデルの性能を比較することができる(334, 335, 419)。既存の予測因子の組み合わせに加えたときの、特定の予測因子による増分の評価のためにNRIが算出されている際には(item 10dにある注意書きを参照)、単一のsummary NRIに加えて、NRIのイベント分と非イベント分をそれぞれ分けて両者を報告すべきである(351, 357, 420, 421)。

純便益や相対有用値のような決定分析指標は通常単一の推定値よりも視覚的に表現される(361-363, 422)。これらのグラフのX軸は、生検をする患者の癌の最小確率や医師がひとつの心血管イベントを予防するのに薬剤で治療を行いたい患者の人数などの患者や臨床医の好みの指標である(117)。X軸の範囲は一般的に実臨床での合理的な変化を表現するよう選択される。例えば、癌の可能性が75%あれば患者は生検を考慮しないだろうと仮定したのであれば、前立腺癌の生検の閾値として80%を含めることの正当性はほとんどない。

純便益であるy軸は、偽陰性の結果と比較した偽陽性のコストを与える因子で重み付けした、真の陽性結果と偽陽性の結果の数の差である。例えば(Figure 18)、特定の閾値で比較した2つのモデルが純便益0.005の差があるとき(つまり、モデルA(QRISK2-2011) -モデルB(NICE フラミンガム))、これは真の陽性所見の純増、つまり偽陽性所見の増加無しにモデルAを用いることで1000人あたりで5人の真の陽性アウトカムを識別した、と解釈する。決定解析尺度の視覚的表現では、　負の純便益にその大半を充てるのは避けるべきである。曲線は平滑にすべきである、つまりもしサンプルサイズが適度にあれば、研究者は統計的な平滑化を用いるか、純便益を更に広い間隔で計算することができる(例: 5%ごと)。

### モデルの更新

### Item17. もしなされているのであれば、モデルのアップデートからの結果（モデルの仕様、モデルの性能）を報告する。[V]

**例**

再較正モデルにおいて、全ての回帰係数は較正モデルの傾きで掛け算されている(男性0.65、女性0.63)。切片は元の値に較正の傾きをかけて較正モデルの切片を加えて調整されている(男性-0.66、女性-0.36)。改定モデルを作成するため、再較正モデルの変数を加えた予測因子の回帰係数をさらに調整した。男性は、回帰係数は、以前の訪問の延期、以前の訪問からの時間、⊿Hb、季節といった予後因子でさらに調整されている。女性は、ヘモグロビンの減少のリスクを計算するための再較正と改訂をしたモデルの正確な公式のために、回帰係数は、前回訪問における減少や⊿Hbなど、この論文のオンライン版において入手可能なサポート情報で更に調整された(370)。[診断; モデル更新; ロジスティック回帰]

アプローチ１における誤較正{mis-calibration}は再較正の必要性を指し示しており、アプローチ2でlogit(P(Y = 1)) = a + b\*logit(p)を当てはめるときに我々は均一シュリンケージ因子(uniform shrinkage factor)を得ている。我々は推定値a=1.20, b=0.11を得ており、重シュリンケージを示している(368)。 [診断; モデル更新; ロジスティック回帰]

オリジナルの予測モデルの性能を、lasso法で選択された遺伝学的変数によって拡張された別のモデルと比較した結果をTable3に示す。尤度比テストが2つのモデル間で適合度検定のために実施されている。AUC曲線(訳注: AUCのCはcurveなので「曲線」は不要と考えられ、原文の誤りと思われる)はオリジナルのモデルで0.856であった。予測モデルへのTLR4 SNPsの追加はAUCをわずかに減少させた。TLR9-1237の追加はAUCをわずかに0.861へ増加させたが、有意ではなかった(p=0.570)。NOD2SNPsはモデルを改善させなかった(423)。[予後; モデル更新; ロジスティック回帰]

**解説**

既存モデルに新たなデータを加えた時の性能はしばしばオリジナルのモデルより性能が落ちることがある。そして研究者はいくつかの方法でモデルを更新するか、再較正するかを決められる(Table 3 とitem 10e)。もし、検証研究の結果に基づいて予測モデルを更新するのであれば、著者は更新されたモデルのすべての側面について報告すべきである。更新の方法によって(Table 3)、報告内容には、再推定した切片の報告・更新された回帰係数(例えば、検証セットにおけるオリジナルモデルの較正曲線の傾きを使用する)・新しい予測因子を含めた推定されたモデルの回帰係数、が含まれる。モデルの更新はCox回帰モデルの文脈においては更に複雑である(309,373)。更新されたモデルは本質的に新しいモデルである。故に、更新されたモデルは、読者が次の妥当性研究や実臨床で、個々の患者の予測ができるように十分な詳細を報告すべきである(item 15aと15b)。更新されたモデルの性能も報告されるべきである(item 16)

## 考察

### 限界

### Item 18. 研究の限界（サンプルに代表性がないこと、予測因子に対してイベントが少ないこと、欠測値など）について議論する。[D; V]

**例**

ICU滞在の長期化に関する予測モデルのもっとも重要な限界はその複雑性にある。ICUの長期滞在に関する因子の膨大さがこの複雑性を反映していると考える。この複雑性により、自動化されたデータ収集と計算が本質的に要求されている。最近では多くの病院で、最新の健康情報技術を簡単に使えないことが、モデルの広い応用への大きな障壁の一つとなっている。より多くの施設で業務フローに電子カルテがくみいれられるほど、ここで述べられているようなモデルは価値が大きくなるだろう。結果にはさらにいくつかの限界がある。第一に、各国間の差異はICU滞在に影響することから、モデルの有効性は米国に限定されること。ICU滞在におけるこれらの差異はICU滞在5日という長期化に関する閾値に不利に影響しやすい。第二に、初日の身体的情報は合併症や治療反応性を説明するのに早すぎる一方で、5日目の情報もなおこれらの効果の説明としては早い。既存の研究では半数以上の合併症がICU治療の5日目以降に発生していることを示している。第三に、この複雑性にも関わらず、モデルは他のICU滞在延長の既知の因子について説明できない。これらの因子には院内感染やDNR (Do not resuscitate)オーダー、ICU専属医、ICUで発生した麻痺、ICUでの鎮静が含まれる。第四に、モデルの最大の不正確さはICU滞在2日以下の患者に関する予測能の悪さである。これらの説明としては、夜間退院や週末の移送の回避やICU滞在6日目から8日目の合併症発生の頻発によると考えられる(424)。[予後; 開発; 検証]

この論文にはいくつかの限界がある。第一に、ひとつの専門科の一プログラムでの研修医のパフォーマンスの評価を代表していることである。加えて我々のプログラムは米国の医学生全体の一部をみているに過ぎない。我々の知見の他のセッティングやプログラムにおける再現性は不明である。第二に、我々は研修医のパフォーマンスを評価するのに累積的な評価と同時に主観的・全体評価を使用した。評価者間信頼性は高かったが、臨床評価のゴールドスタンダードはなく、臨床のパフォーマンスを評価する最良の方法については議論の余地がある。最後に回帰分析のr2=0.22は平均の性能評価の分散の多くが説明つかないことを示している。これは、レジデンシー応募におけるリーダーシップやチームワーク、プロフェッショナリズムのような重要な領域の情報が限られていることによるかもしれない(425)。[予後; 開発]

**解説**

最良の実施をされた予測モデルの研究でさえ、考慮すべき多くの限界を有する可能性が高い。しかし、最も影響力のある雑誌であっても、掲載されている多くの論文で限界についての報告がされていない(426)。更に、研究デザインや知見に起因する注意への適切な考察が欠けた、結果の拡大解釈は分子分類の研究ではよくみられる(158)。

出版後に問われた時、多くの共著者は印刷されたディスカッションは彼らの意見全てを表現していないと感じるし、彼らはしばしば、出版されたディスカッションの限界や注意の列挙が短いことをつけ加える(427)。しかし、限界への明確な認識は科学的な著作の大事な側面の一つであり、科学論文のディスカッションで最も価値のある部分であると繰り返し述べられている(428, 429)。限界への認識は研究を弱めることはなくむしろ強化する。限界は展望に置き換えられる必要があり、研究結果に影響するそれぞれ別々の問題として特徴づける努力がされるべきである。いくつかの場合において、その影響ははっきりとは分からず、実際には推測することは不可能である一方で、バイアスの方向を安全に予測でき、影響を実際的な確実性をもって評価できる場合もある。

限界は、研究デザインや研究の実施や解析のどのような局面とも関連することがある。これには、以下に限るものではないが、研究対象集団のタイプ、対象者の選定(代表性)、予測因子の選択、定義の頑健さ、予測因子とアウトカム両者のデータ収集に使用された方法、サンプルサイズ(特に、複雑性と予測因子とアウトカムのイベントの数の比較において)、フォローアップと確認方法の長さ、解析の多重度、欠測値、過剰適合、内的検証の過程、開発コホートと検証コホートの違い(適用できるのであれば)が含まれる(430)。限界がモデル開発や検証のいずれかまたは両者に影響しているか、多変量モデルの信頼性、適用性、一般化可能性へどのような影響が予測されるかについて議論すべきである。

例えば、既知の関連の強い予測因子の脱落は明確に認めるべきであるし、これらの既知の予測因子はリストを作成するべきである。もし可能であれば、脱落している予測因子が将来の研究や実用化において考慮する必要があるかどうかや、包含されている予測因子が既に十分に脱落した因子から情報を得ているか、を明らかにするのに有用かもしれない。他の例では、もし過剰適合が疑われるのであれば、これも認めるべきであり、どれほど重大な問題が予測されるのか、モデルの性能はどの程度過大評価されているか、これらの懸念を緩和させるようないくつかの更なる検証やモデルの更新(再較正を含む; item 10eと17)、継続的な改善戦略(例; QRISK2(117,431-433)を待つのに対して、この過大評価を実臨床モデルに適応する決定に影響させるかどうかについて、いくつかのステートメントに記述されるべきである.

異なる集団での検証を行っていない単一集団でのモデル開発に関する論文は、他の限界もあり得ることに加えて、更に外的検証を行ってないことを主要な限界として必ず述べるべきである

### 解釈

### Item 19a. 検証を行う場合、結果を考察する際に、開発データでの性能、およびその他の検証データを言及する。[V]

**例**

ABCD2スコアは、脳卒中のハイリスクの所見を有した患者の2つの異なるデータセットを統合したJohnstonとRothwellによって率いられたチームの努力の結集である。Rothwellのデータセットは小さく、プライマリ・ケア医を受診した患者に由来し、1～3日後に神経内科医に予測変数を評価されたデータセットである。Johnsonのデータセットはカルフォルニアの一過性脳虚血発作の患者の後方視研究に由来する。ABCD2 scoreを評価したその後の研究には後方視研究やデータベース研究がある。Ongらはスコア2点以上で高リスクと評価してから7日以内の脳卒中発症の感度が96.6%であると示したが、83.6%の患者が高リスクと分類された。Fothergillらは後方視的に284名を解析し、4点以下をカットオフとした7日以内の脳卒中患者のうち36名中4名を見逃したと報告している。Asimosらは後方視的にABCD2スコアをデータベースから算出したが、37%の患者はスコアを算出できず、その中には7日以内に脳卒中を起こした373名のうち154名の患者を含んでいた。SheehanらはABCD2スコアで一過性脳虚血発作や軽微な脳卒中と他の疾患による一過性の神経徴候を良く識別することを認めているが、彼らは引き続き起こる脳卒中に対するスコアの予測能の正確性について評価しなかった。Tsivgooulisらは、一過性脳虚血発作で入院した患者の前向き研究のデータを元に、ABCD2スコア2点以上をハイリスクとして使用することを支持している。GilesとRothwellによるシステマティック・レビューではsearch criteriaを満たす全ての研究によるプールされた AUCが0.72 (95%CI0.63-0.82)であり、もとの作成研究を除外した後のAUCが0.69 (95%CI0.64-0.74)であることを示した。我々の研究のAUCは、これらの研究の信頼区間の下限値にあり、0.5に近かった。[予後]

**解説**

既存モデルの検証を示すときには、検証されたモデルが開発されたものと同等であるかどうかや、何らかの差異があるかどうか、もしそうならそれは何故かについて明確に述べるべきである(item 12)。検証研究におけるモデルの性能は、元の開発研究と既存の他の検証研究の文脈を踏まえて議論するべきである。比較に影響しうるあらゆるバイアス同様に、主要な結果を強調すべきである。

検証研究が異なる性能を示す場合(通常、悪い性能を示す)、解釈を強化するためにその理由について議論すべきである。例えば、理由には診断関連群や予後因子やアウトカムの定義や測定、(当てはまるのであれば)追跡期間の違いが含まれる。1つ以上のモデルが単一のデータセットで検証される、いわゆる競合検証{comparative validation}のときは、主要な結果は、対照に影響するバイアスを含めて強調するべきである(47, 48)。

### Item 19b. 目的、限界、似た研究からの結果、他の重要なエビデンスを踏まえて、全ての結果の解釈を提示する。[D; V]

**例**

我々のモデルは将来の腎不全を予測するため、人口統計属性と慢性腎臓病の重症度のマーカーを使用している。Kaiser PermanenteとRENAAL studyグループら以前の研究者同様に、我々は推定GFR(糸球体濾過量)の低値、高アルブミン尿、若年者、男性が腎不全の進行が速いことを予測することを認めた。加えて、低アルブミン血症、低カルシウム血症、低重炭酸血症と高リン血症も腎不全の高リスクを予想し、推定GFRとアルブミン尿の予測能に加えた。

これらのマーカーは糸球体濾過量をよく推定でき、尿細管機能障害や潜在的な炎症や栄養不良を反映し得る。これらの検査データは以前からCKDの進行に関連していたが、我々はこれらをひとつの式に統合した(リスク計算とTable 5、そしてスマートフォンアプリがwww.qxmd.com/Kidney-Failure-Risk-Equationで入手可能)。加えて我々は病歴(糖尿病・高血圧の状態)や身体所見(収縮期血圧、拡張期血圧、体重)から得られる変数を加えてもモデルの性能が変わらないことを示している。これら他の変数はCKDの診断や管理に明らかに重要だが、モデルの性能の改善がないことはこれらの状況がCKDでは頻度が高いこと、推定GFRとアルブミン尿を既に考慮にいれた後の病気の重症度に対しての精度が低いことを反映しているかもしれない(261)。[予後; 開発; 検証]

**解説**

研究結果の解釈とは、発見されたことを他のエビデンスの文脈内に位置づけることである。他のエビデンスには、同じ多変量モデルや、同じ、もしくは似たアウトカムについての異なるモデル、関連のありそうな他のタイプのエビデンスについての類似した先行研究が含まれるだろう。しばしば、同じ、もしくは似た目的を持つ多くの予測モデルがある。例えば、死亡率のみについて、1年間で118の異なる予測ツールの240の評価研究が出版されている (65)。

同じ標的集団やアウトカムのための多くの使用可能な予測モデルがあるときには、先行の開発されたモデルに対するモデルの系統的な並列{ juxtaposition }が新しいモデルの利点と欠点を見分けるのに有用かもしれない。このような比較はレビューができるのであれば、理想的には先行研究のシステマティック・レビューによることになると考える(47, 48, 435)。そうでなければ、著者は、エビデンスの基礎の情報提供と更なる検証研究や実臨床への適応へのアクションプランの観点で、最低限の先行のエビデンスの非公式なレビューを行うことを考慮し、現在の研究に競合する主な研究に関する議論を行うことが必要である。モデル構築、予測因子の考慮、適応集団とセッティング、性能の違いと、検証の過程の強みに関する解説は有用かもしれない。追加で考慮することとしては、結果の解釈を支えることであり、以下の内容を含む。研究の限界(item 18で議論した通り)。当初の研究目的と一致するかどうか、もしそうでなければなぜか。そして異なるセッティングでのモデルの実行可能性の観点、およびモデルを導入することが、どのように他の実臨床に適応するか、変化をもたらすかといったことである。

場合によっては、他の関連するエビデンスについて考察することも面白いだろう。例えば、モデルに含まれる予測因子の生物学的な蓋然性について有用なデータや、モデルの中のいくつかのデータが部分的に重要かについての示唆を与えるようなその他のデータがあるかもしれない。経験的証拠として、モデルにある特定の予測因子を含むことを支持する生物学的妥当性に関るエビデンスの示し方について、研究者は全く体系的でないことが示唆されている(436)。バランスのとれた視点を持つこと、支持的・反証的エビデンスの両者について存在の有無にかかわらず、議論する労力を払うべきである。

### 見込み

### Item 20. モデルの潜在的な臨床応用および今後の研究の見込みについて議論する。 [D; V]

**例**

インフルエンザ罹患の可能性については、そのコミュニティーでのインフルエンザ発症確率、診察の結果、（任意で）インフルエンザの迅速検査の結果に依存する。Centers for Disease Control and Preventionからのデータにて季節ごとのインフルエンザ発症確率を手に入れることにした。最近のsystematic reviewにて、迅速検査は、季節性インフルエンザにておよそ感度72%、特異度96%であったとの報告があった。季節ごとのインフルエンザの発症確率や迅速検査の精度を用いて、インフルエンザスコアの尤度比を1、検査をするかしないかの閾値を10%、検査をするか治療をするかの閾値を50%として、インフルエンザが疑われる患者の評価方法の推奨を Table 5に要約した。抗インフルエンザ薬を使用制限したいと考える医師は、インフルエンザが流行っている季節の高リスク患者であっても迅速テストを考慮すべきである。経験的な治療は、合併症発症の高リスク患者で考慮されるかも知れない(181)。[診断; 開発; 検証; 臨床使用での見込み]

これらの結果をより深く理解するために、いくつか重要な点を示す必要がある。

まず、データの元となっている研究は外来患者も含まれているが、外来患者でのPONV[postoperative nausea and vomiting]の発症割合は実質的には高くない(34%)、また外来患者では手術のタイプも異なる(例、腹部手術は行わない)ので、今回の解析では我々は意識して入院患者のみに限定した。そのため、我々の結果はまずは入院患者に適応されるべきである。現在のところ、外来患者および入院患者両方から得られた法則は利用できないことに注意すべきである。特に外来手術が増えてくることから、これはなお今後の研究のテーマである(437) 。[予後; 増分価値; 臨床使用での見込み]

我々の研究は認めなければならないいくつかの欠点がある。インフルエンザ流行シーズンや季節ごとのインフルエンザの事前確率が異なる2つの国の患者を組み込んでいるので、一般化可能性が高まったデータにはなったが、組み入れ基準がやや異なる2つのグループからのデータを統合して利用したことが挙げられる。また、データ収集は成人に限定しているため、これらの結果を若い患者に適応できるかは不明である。シンプルではあるが、点数を計算することが複雑すぎるために覚えるのが難しい可能性があり、スマートフォンやインターネットを使用したアプリケーションのようなプログラムにより使用の手助けができると思われる(181)。 [診断; 開発; 検証; 限界; 研究での見込み]

**解説**

考察の項では、著者がさまざまなレベルでその結果の見込みを議論する機会および責任の両者がある。予測モデルは、その使用方法にいろいろな可能性がある。Item 3a（モデルための背景と意義）にて、研究者はそのモデルに関する可能性について記述することを求められている。考察は、その研究での知見について潜在的な臨床応用を議論する部分である。明らかなこととして、新しく開発されたモデルでは、検証研究を行うことが論理的に考えられる次のステップである。そのため、実際に臨床でどのように使用するか改まった議論をすることはより難しいと思われる。実際に、最初の開発研究のみに基づいたモデルの適応を推奨することは、防がねばならない。

同様に、臨床ガイドラインでも検証がなされていない予測モデルの使用に関して推奨すべきではない。むしろ、臨床では、利用可能性と、他の対象者データでのモデルの精度のエビデンスの統合、つまりモデルの転用可能性に基づいて臨床での推奨がなされるべきである。

前向きにデザインされた研究であったとしても外的検証研究は、あるモデルの結果が診療上の意思決定に影響したり、健康関連のアウトカムを改善する程度を示すことはできないことについて強調しておく。

意思決定、医師の習慣、患者アウトカムへの効果は、単一のコホート研究やモデル検証研究(20, 28, 33)ではなく、比較研究(可能な限りランダム化［438-440］)によってのみ評価可能である。不幸にも、モデルの影響研究は言うまでもなく、外的検証研究はまれである (441, 442)。

「どのような患者に結果を適応するか」という観点から、予測が評価されたセッティング（プライマリ・ケア、病院）、地理的状況、年齢、性別、医学的問題の臨床的特徴を含めて議論されるべきであろう。2番目の項目は実際にどのようにその（予測）ルールが適応されるかに関わる。例えば、検証された診断モデルの目的は診断確定であるのか除外であるのか、予測モデルのカットオフ値はそれぞれの目的に対してどの値であるのか、結果にどんな可能性があるのか(例えば、さらなる精密検査、偽陽性、偽陰性)である。

短期的な可能性に関する見込みを超え、さらなる研究に関する具体的な示唆が、現在の研究の限界にもとづき行われる。以下のような事柄である、新規に開発されたモデルの他のデータセットでの検証の必要性や、そのゴール(と他の予測因子の潜在的な有用性)を完成するためのルールの力、診療の助けとなる閾値の選択、臨床適応での問題。

## 他の情報

### 補足情報

### Item 21. 研究プロトコルやウェブ計算機、データセットといった補足情報の利用可能性について情報を提示する。[D; V]

**例**

RISK−PCI trialでのデザインや方法は以前に公表されている [参照]。

簡単にいえば、RISK−PCIは、もともとクロピドグレル600mgを内服していた患者が経皮的冠動脈形成術を受けた際に大きな心臓血管イベントを発症することの予測モデル作成および妥当性を検証するためにデザインされた観察的、縦断的、コホート、単施設での研究である。患者は、2006年2月から2009年12月まで登録された。

それぞれの患者からインフォームドコンセントを得た。

研究プロトコルは、ヘルシンキ宣言に従った。

施設の倫理委員会にてより承認され、the Current Controlled Trials Register　- SRCTN83474650- (www.controlled-trials.com/ISRCTN83474650) に登録した (443)。[予後; 開発]

男性および女性用のReynolds Risk Scoresのための使用者が扱いやすい計算機はwww.reynoldsriskscore.org からアクセスが無料でできる(44) [予後; 増分価値]

QFracture- Scores 算出のための開示された情報コードがwww.qfracture.orgよりGNU lesser general public licence-version 3で利用できる（315）[予後; 検証]

**解説**

人を対象としたすべての研究はプロトコルに基づくべきである(445, 446)。予測モデル研究のプロトコルは明確な研究目的、研究デザイン、予測因子とアウトカムの記述、統計解析計画と続く。予測モデルの開発および妥当性検討のための研究は、解析が実行される前に、詳細なプロトコルを注意深く作成することによってより良いものとなるであろう。これらのプロトコルはときどき出版される(447-464)。もし出版されていれば、計画されていたことと実際に行われたことの比較が読者に可能となる。もし出版しないのであれば、雑誌に投稿する際にプロトコルを一緒に送るようにし、査読者が出版された論文を評価する際に役立つように、可能であればオンラインでの付録として出版された論文に添付するように、我々は著者に促す。

普段の臨床で、もしくは、さらなる研究にて予測モデルを使用するために、個人の予測確率を得られるようにし、研究者が予測モデルの検証と更新をできるように十分に詳細な情報(items 15a, 15b,と16)を論文に記載することを必要とする。さらに、開発された(例、www.outcomes-umassmed.org/GRACE/) ウェブ上の計算機や独立型のアプリケーション(例、タブレットのような電子機器)をどのように使うかについて、著者は詳細を示すべきである。（モデルが）とても複雑で論文に詳細まで充分に記述できないようなまれな状況、あるいはモデルが継続的に更新される(例えば、QRISK2 [139])場合は、予測因子の計算のためのプログラムコードにどこでフルアクセスできるのか詳細を示すべきである。

データセットやコンピューターコードは解析の再現のためだけでなく、個々人の対象者データによるメタアナリシスのためににも（468-473）、可能な限り公開するべきと認識されるようになってきている（27, 465-467）。他の研究者に生データを準備し共有するためのガイダンスが作成されており、著者の助けとなるであろう(271)。Marchionniらによる模範となる論文において、予測モデル開発のすべての過程について透明性をもって再現性を確認するために必要なテンプレートの原型が提供されている(474)。可能であれば、データ解析に必要な情報コードにアクセスできる詳細を提供すべきである。

現在、観察研究での登録の必要性はなく、登録に賛成する意見があった(475-478)が、反対の意見もあった(479-481)。ClinicalTrials.govを含む多くの臨床研究のレジストリーでは観察研究も登録可能であると明確にのべている(482)。いくつかのタイプの観察研究では、詳細を前もって計画することが難しいが、予測モデルの作成および妥当性確認を目的とした新たな患者を集積する（前向きな）予測モデル研究では、この考えは適応されない(476)。

### 資金源

### Item 22.研究における研究資金源および資金提供者の役割について提示する。[D; V]

**例**

Reynolds Risk Scoreプロジェクトは Donald W. Reynolds Foundation (Las Vegas, Nev)からの研究者主導研究資金によりサポートを受け、さらにthe Doris Duke Charitable Foundation (New York, NY)と the Leducq Foundation (Paris, France)から補助を受けた。Women's Health Study cohortはthe National Heart, Lung, and Blood Institute and the National Cancer Institute (Bethesda, Md)からの研究資金を受けた (208)。[予後; 開発]

Weill Cornell Medical CollegeのThe Clinical and Translational Service Center からデータ解析の一部でサポートを受けた。資金提供者による解析のデザイン、その解釈、出版原稿の提出への一切の関与はなかった(380)。[診断; 開発; 検証]

**解説**

予測モデル研究では、前向きな研究であってさえも、ほとんど、まったく資金提供を受けない傾向にあり、そのことが多くの研究のクオリティの低さに寄与していることが示唆されている。通常であれば資金獲得を考えるような、研究計画段階に多くはピアレビューがないままプロトコルが作成されている(472)。

著者は、研究を作成するために受けた資金提供のすべての情報を明らかにすべきであり、その提供者の研究デザイン、実施、解析、報告での役割を述べるべきである。

もし、資金提供者がいない場合、著者はそのように述べるべきである。同様に、外部資金提供がない場合も、そのように明確に述べるべきである。

ガイドラインに組みこまれるモデルでは、その予後モデル開発に関わるメンバーだけでなく、ガイドラインに作成に関連するすべてのメンバーで可能性のある資金提供と利益相反を示すべきである(316, 483, 484)。

# 結語

予測モデルを扱った研究は豊富で、モデルの開発や検証、更新、拡張など、次々と出版されとどまる気配がない。TRIPOD声明は、更新の有無に関わらず１つもしくはそれ以上の診断および予後予測モデルの開発、妥当性検討の研究を報告するための手引きである。完全で、透明性のある報告のみが研究の強みと弱点を明らかにできる。そうすることで、その解釈が容易になり、利用することができるようになる(485-487)。完全な報告は、今後の予測モデルの土台となり、特に研究者が検討や現存する予測モデルとの比較を行うことを可能とする。完全な報告は、日常診療での使用でも妥当性のある予測モデルの理解と解釈に役立てることもできる。TRIPOD声明は、査読者や雑誌編集者による予測モデル研究の評価をおこなう際に役立つであろう。TRIPODは予測モデル研究でのデザイン、実施、解析でも助けとなりうる。

TRIPODは多くの専門領域にわたる24人の専門家によって作成され、そのうちには 報告ガイドラインであるCONSORT (96), STROBE (97, 99), PRISMA(488), REMARK (98), GRIPS (101), STREGA(489), STARD (100), ARRIVE (490), and CARE (491)に関わっていた者もいる。専門的な知識を使った合意に基づくガイドライン作成の集約的な経験を生かしているため、我々は、報告ガイドラインを作成する際の手引きに忠実であった

(113)。チェックリストのそれぞれのitemについて、我々は広く議論し、意義を示し、良い報告から役立つ例を記載した。可能な限り、出版されている総説から関連性のある経験的証拠を参照した。さらに、我々は、予測モデルの開発や検証の中で重要な項目に関してさらなる議論をBoxを使って示した。

TRIPODが著者、査読者、雑誌編集者の負担を増やしているという指摘があるかも知れない。しかし、我々はTRIPODを使用することがおそらく査読時間の減少になり、訂正依頼を減らし、適正な査読過程を確実にするのに役立つと考える。チェックリストに含まれたitemは、合意に至るまでの多くの議論を反映し、十分な情報にもとづく研究の質の評価、バイアスリスクと臨床的な関連性を報告することを可能とし、結果を使用可能とする最低限のセットによって構成されている(532)。

報告ガイドラインは、誤って研究の創造性を抑えてしまうことも考えられる。他の報告ガイドラインと同様に、TRIPODではどのように解析を実行するのかを求めておらず、むしろ、適切な情報が報告されることを確実にすることを目的としている。

最後に、TRIPOD声明は、予測モデル研究の方法論が発展を続けるように、継続的な評価を必要とし、必要に応じて洗練される発展性のある文書としてとらえるべきである。

TRIPODのウェブサイト(www.tripod-statement.org)では、議論のためのフォーラム、チェックリストとこの解説と詳細文章を改善するための示唆、予測モデル研究に関連したリソースを提供している。我々は、チェックリストの翻訳を促進し、それをウェブサイト上で使用できるようにすることも構想している。TRIPODに関連する告知や情報はTRIPOD Twitter アドレス(@TRIPODStatement).を通して伝える予定である。TRIPODは健康に関する研究報告のためのEQUATOR Network library ([www.equator-network.org](http://www.equator-network.org/))にもリンクし広報する予定である。

# 引用文献

1. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009; 338:b375. [PMID: 19237405]

2. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.

3. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985; 313:793-9. [PMID: 3897864]

4. Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ. 2011; 343:d5888. [PMID: 21968126]

5. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol. 2006;6:18. [PMID: 16613605]

6. Kattan MW, Vickers AJ. Incorporating predictions of individual patient risk in clinical trials. Urol Oncol. 2004;22:348-52. [PMID: 15283895]

7. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA. 2007;298:1209-12. [PMID: 17848656]

8. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al; PROGRESS Group. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med. 2013;10: e1001380. [PMID: 23393429]

9. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al; PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10: e1001381. [PMID: 23393430]

10. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ. 2009; 338:b604. [PMID: 19336487]

11. Collins GS, Altman DG. Identifying patients with undetected renal tract cancer in primary care: an independent and external validation of QCancer® (Renal) prediction model. Cancer Epidemiol. 2013; 37:115-20. [PMID: 23280341]

12. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87. [PMID: 8668867]

13. Canet J, Gallart L, Gomar C, Paluzie G, Valle` s J, Castillo J, et al; ARISCAT Group. Prediction of postoperative pulmonary complications in a population-based surgical cohort. Anesthesiology. 2010; 113:1338-50. [PMID: 21045639]

14. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. Eur J Cardiothorac Surg. 2012;41:734-44. [PMID: 22378855]

15. Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Mo¨ hlig M, et al. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care. 2007;30:510-5. [PMID: 17327313]

16. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. BMJ. 2009;338:b880. [PMID: 19297312]

17. D’Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117:743-53. [PMID: 18212285]

18. North RA, McCowan LM, Dekker GA, Poston L, Chan EH, Stewart AW, et al. Clinical risk prediction for pre-eclampsia in nulliparous women: development of model in international prospective cohort. BMJ. 2011;342:d1875. [PMID: 21474517]

19. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338: b605. [PMID: 19477892]

20. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691-8. [PMID: 22397946]

21. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol. 2008;61:1085-94. [PMID: 19208371]

22. Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. Eur J Clin Invest. 2012; 42:216-28. [PMID: 21726217]

23. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003; 56:441-7. [PMID: 12812818]

24. Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000; 19:1059-79. [PMID: 10790680]

25. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making. 2001;21:45-56. [PMID: 11206946]

26. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19:453-73. [PMID: 10694730]

27. Ioannidis JPA, Khoury MJ. Improving validation practices in “omics” research. Science. 2011;334:1230-2. [PMID: 22144616]

28. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130:515-24. [PMID: 10075620]

29. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. JAMA. 2000;284:79-84. [PMID: 10872017]

30. Taylor JM, Ankesrt DP, Andridge RR. Validation of biomarkerbased risk prediction models. Clin Cancer Res. 2008;14:5977-83. [PMID: 18829476]

31. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol. 2008;61:76-86. [PMID: 18083464]

32. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774-81. [PMID: 11470385]

33. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144:201-9. [PMID: 16461965]

34. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9:1-12. [PMID: 22629234]

35. Rabar S, Lau R, O’Flynn N, Li L, Barry P; Guideline Development Group. Risk assessment of fragility fractures: summary of NICE guidance. BMJ. 2012;345:e3698. [PMID: 22875946]

36. National Institute for Health and Care Excellence. Lipid modification: cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. Clinical guideline CG67. London: National Institute for Health and Care Excellence; 2008. Accessed at http://guidance.nice.org.uk/CG67 on 30 October 2011.

37. National Osteoporosis Foundation. Clinician's guide to prevention and treatment of osteoporosis. Washington DC: National Osteoporsis Foundation; 2010. Accessed at http://nof.org/hcp/clinicians-guide on 17 January 2013.

38. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III) final report. Circulation. 2002;106:3143-421. [PMID: 12485966]

39. Goldstein LB, Adams R, Alberts MJ, Appel LJ, Brass LM, Bushnell CD, et al; American Heart Association; American Stroke Association Stroke Council. Primary prevention of ischemic stroke: a guideline from the American Heart Association/American Stroke Association Stroke Council: cosponsored by the Atherosclerotic Peripheral Vascular Disease Interdisciplinary Working Group; Cardiovascular Nursing Council; Clinical Cardiology Council; Nutrition, Physical Activity, and Metabolism Council; and the Quality of Care and Outcomes Research Interdisciplinary Working Group. Circulation. 2006;113: e873-923. [PMID: 16785347]

40. Lackland DT, Elkind MS, D’Agostino R Sr, Dhamoon MS, Goff DC Jr, Higashida RT, et al; American Heart Association Stroke Council; Council on Epidemiology and Prevention; Council on Cardiovascular Radiology and Intervention; Council on Cardiovascular Nursing; Council on Peripheral Vascular Disease; Council on Quality of Care and Outcomes Research. Inclusion of stroke in cardiovascular risk prediction instruments: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2012;43:1998-2027. [PMID: 22627990]

41. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak. 2006;6:38. [PMID: 17105661]

42. Shariat SF, Karakiewicz PI, Margulis V, Kattan MW. Inventory of prostate cancer predictive tools. Curr Opin Urol. 2008;18:279-96. [PMID: 18382238]

43. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest. 2009;27:235-43. [PMID: 19291527]

44. van Dieren S, Beulens JW, Kengne AP, Peelen LM, Rutten GE, Woodward M, et al. Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review. Heart. 2012;98:360-9. [PMID: 22184101]

45. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011;9:103. [PMID: 21902820]

46. Ettema RG, Peelen LM, Schuurmans MJ, Nierich AP, Kalkman CJ, Moons KG. Prediction models for prolonged intensive care unit stay after cardiac surgery: systematic review and validation study. Circulation. 2010;122:682-9. [PMID: 20679549]

47. Collins GS, Moons KG. Comparing risk prediction models. BMJ. 2012;344:e3186. [PMID: 22628131]

48. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012;344:e3318. [PMID: 22628003]

49. Seel RT, Steyerberg EW, Malec JF, Sherer M, Macciocchi SN. Developing and evaluating prediction models in rehabilitation populations. Arch Phys Med Rehabil. 2012;93(8 Suppl):S138-53. [PMID: 22840880]

50. Green SM, Schriger DL, Yealy DM. Methodologic standards for interpreting clinical decision rules in emergency medicine: 2014 update. Ann Emerg Med. 2014;64:286-91. [PMID: 24530108]

51. Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. Ann Intern Med. 2007;146:450-3. [PMID: 17339612]

52. Groves T, Godlee F. Open science and reproducible research. BMJ. 2012;344:e4383. [PMID: 22736475]

53. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney were poorly reported and often developed using inappropriate methods. J Clin Epidemiol. 2013;66:268-77. [PMID: 23116690]

54. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med. 2010;8:20. [PMID: 20353578]

55. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. BMC Med. 2010;8:21. [PMID: 20353579]

56. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer. 2004;91:4-8. [PMID: 15188004]

57. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med. 1993;118:201-10. [PMID: 8417638]

58. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA. 1997;277:488-94. [PMID: 9020274]

59. Steurer J, Haller C, Ha¨uselmann H, Brunner F, Bachmann LM. Clinical value of prognostic instruments to identify patients with an increased risk for osteoporotic fractures: systematic review. PLoS One. 2011;6:e19994. [PMID: 21625596]

60. van Dijk WD, Bemt L, Haak-Rongen S, Bischoff E, Weel C, Veen JC, et al. Multidimensional prognostic indices for use in COPD patient care. A systematic review. Respir Res. 2011;12:151. [PMID: 22082049]

61. Hayden JA, Coˆ te´ P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med. 2006;144: 427-37. [PMID: 16549855]

62. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat. 2012;132:365-77. [PMID: 22037780]

63. Mushkudiani NA, Hukkelhoven CW, Herna´ndez AV, Murray GD, Choi SC, Maas AI, et al. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. J Clin Epidemiol. 2008;61:331-43. [PMID: 18313557]

64. Rehn M, Perel P, Blackhall K, Lossius HM. Prognostic models for the early care of trauma patients: a systematic review. Scand J Trauma Resusc Emerg Med. 2011;19:17. [PMID: 21418599]

65. Siontis GC, Tzoulaki I, Ioannidis JP. Predicting death: an empirical evaluation of predictive tools for mortality. Arch Intern Med. 2011;171:1721-6. [PMID: 21788535]

66. Medlock S, Ravelli ACJ, Tamminga P, Mol BW, Abu-Hanna A. Prediction of mortality in very premature infants: a systematic review of prediction models. PLoS One. 2011;6:e23441. [PMID: 21931598]

67. Maguire JL, Kulik DM, Laupacis A, Kuppermann N, Uleryk EM, Parkin PC. Clinical prediction rules for children: a systematic review. Pediatrics. 2011;128:e666-77. [PMID: 21859912]

68. Kulik DM, Uleryk EM, Maguire JL. Does this child have appendicitis? A systematic review of clinical prediction rules for children with acute abdominal pain. J Clin Epidemiol. 2013;66:95-104. [PMID: 23177898]

69. Kulik DM, Uleryk EM, Maguire JL. Does this child have bacterial meningitis? A systematic review of clinical prediction rules for children with suspected bacterial meningitis. J Emerg Med. 2013;45: 508-19. [PMID: 23910166]

70. Jacob M, Lewsey JD, Sharpin C, Gimson A, Rela M, van der Meulen JH. Systematic review and validation of prognostic models in liver transplantation. Liver Transpl. 2005;11:814-25. [PMID: 15973726]

71. Hussain A, Choukairi F, Dunn K. Predicting survival in thermal injury: a systematic review of methodology of composite prediction models. Burns. 2013;39:835-50. [PMID: 23384617]

72. Haskins R, Rivett DA, Osmotherly PG. Clinical prediction rules in the physiotherapy management of low back pain: a systematic review. Man Ther. 2012;17:9-21. [PMID: 21641849]

73. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. PLoS Med. 2012;9:e1001344. [PMID: 23185136]

74. Echouffo-Tcheugui JB, Batty GD, Kivima¨ki M, Kengne AP. Risk models to predict hypertension: a systematic review. PLoS One. 2013;8:e67370. [PMID: 23861760]

75. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A. Risk prediction models of breast cancer: a systematic review of model performances. Breast Cancer Res Treat. 2012;133:1-10. [PMID: 22076477]

76. van Oort L, van den Berg T, Koes BW, de Vet RH, Anema HJ, Heymans MW, et al. Preliminary state of development of prediction models for primary care physical therapy: a systematic review. J Clin Epidemiol. 2012;65:1257-66. [PMID: 22959592]

77. Tangri N, Kitsios GD, Inker LA, Griffith J, Naimark DM, Walker S, et al. Risk prediction models for patients with chronic kidney disease: a systematic review. Ann Intern Med. 2013;158:596-603. [PMID: 23588748]

78. van Hanegem N, Breijer MC, Opmeer BC, Mol BW, Timmermans A. Prediction models in women with postmenopausal bleeding: a systematic review. Womens Health (Lond Engl). 2012;8:251-62. [PMID: 22554173]

79. Minne L, Ludikhuize J, de Jonge E, de Rooij S, Abu-Hanna A. Prognostic models for predicting mortality in elderly ICU patients: a systematic review. Intensive Care Med. 2011;37:1258-68. [PMID: 21647716]

80. Leushuis E, van der Steeg JW, Steures P, Bossuyt PM, Eijkemans MJ, van der Veen F, et al. Prediction models in reproductive medicine: a critical appraisal. Hum Reprod Update. 2009;15:537-52. [PMID: 19435779]

81. Jaja BN, Cusimano MD, Etminan N, Hanggi D, Hasan D, Ilodigwe D, et al. Clinical prediction models for aneurysmal subarachnoid hemorrhage: a systematic review. Neurocrit Care. 2013;18:143-53. [PMID: 23138544]

82. Wlodzimirow KA, Eslami S, Chamuleau RA, Nieuwoudt M, Abu-Hanna A. Prediction of poor outcome in patients with acute liver failure-systematic review of prediction models. PLoS One. 2012;7: e50952. [PMID: 23272081]

83. Phillips B, Wade R, Stewart LA, Sutton AJ. Systematic review and meta-analysis of the discriminatory performance of risk prediction rules in febrile neutropaenic episodes in children and young people. Eur J Cancer. 2010;46:2950-64. [PMID: 20621468]

84. Rubin KH, Friis-Holmberg T, Hermann AP, Abrahamsen B, Brixen K. Risk assessment tools to identify women with increased risk of osteoporotic fracture: complexity or simplicity? A systematic review. J Bone Miner Res. 2013;28:1701-17. [PMID: 23592255]

85. Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AM, et al. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. BMJ. 2012;345:e5900. [PMID: 22990994]

86. Braband M, Folkestad L, Clausen NG, Knudsen T, Hallas J. Risk scoring systems for adults admitted to the emergency department: a systematic review. Scand J Trauma Resusc Emerg Med. 2010;18:8. [PMID: 20146829]

87. Maguire JL, Boutis K, Uleryk EM, Laupacis A, Parkin PC. Should a head-injured child receive a head CT scan? A systematic review of clinical prediction rules. Pediatrics. 2009;124:e145-54. [PMID: 19564261]

88. Vuong K, McGeechan K, Armstrong BK, Cust AE. Risk prediction models for incident primary cutaneous melanoma: a systematic review. JAMA Dermatol. 2014;150:434-44. [PMID: 24522401]

89. Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data metaanalysis. BMC Med Res Methodol. 2014;14:3. [PMID: 24397587]

90. Huen SC, Parikh CR. Predicting acute kidney injury after cardiac surgery: a systematic review. Ann Thorac Surg. 2012;93:337-41. [PMID: 22186469]

91. Calle P, Cerro L, Valencia J, Jaimes F. Usefulness of severity scores in patients with suspected infection in the emergency department: a systematic review. J Emerg Med. 2012;42:379-91. [PMID: 22142675]

92. Usher-Smith JA, Emery J, Kassianos AP, Walter FM. Risk prediction models for melanoma: a systematic review. Cancer Epidemiol Biomarkers Prev. 2014;23:1450-63. [PMID: 24895414]

93. Warnell I, Chincholkar M, Eccles M. Predicting perioperative mortality after oesophagectomy: a systematic review of performance and methods of multivariate models. Br J Anaesth. 2014 Sep 17. [Epub ahead of print]. [PMID: 25231768]

94. Silverberg N, Gardner AJ, Brubacher J, Panenka W, Li JJ, Iverson GL. Systematic review of multivariable prognostic models for mild traumatic brain injury. J Neurotrauma. 2014 Sep 15. [Epub ahead of print]. [PMID: 25222514]

95. Delebarre M, Macher E, Mazingue F, Martinot A, Dubos F. Which decision rules meet methodological standards in children with febrile neutropenia? Results of a systematic review and analysis. Pediatr Blood Cancer. 2014;61:1786-91. [PMID: 24975886]

96. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332. [PMID: 20332509]

97. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335:806-8. [PMID: 17947786]

98. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM; Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst. 2005;97:1180-4. [PMID: 16106022]

99. Gallo V, Egger M, McCormack V, Farmer PB, Ioannidis JP, Kirsch-Volders M, et al. STrengthening the Reporting of OBservational studies in Epidemiology - Molecular Epidemiology (STROBE-ME): an extension of the STROBE statement. Eur J Clin Invest. 2012;42:1-16. [PMID: 22023344]

100. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Radiology. 2003;226:24-8. [PMID: 12511664]

101. Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ; GRIPS Group. Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. Eur J Clin Invest. 2011;41:1004-9. [PMID: 21434891]

102. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606. [PMID: 19502216]

103. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98:683-90. [PMID: 22397945]

104. Labare` re J, Bertrand R, Fine MJ. How to derive and validate clinical prediction models for use in intensive care medicine. Intensive Care Med. 2014;40:513-27. [PMID: 24570265]

105. Tzoulaki I, Liberopoulos G, Ioannidis JP. Use of reclassification for assessment of improved prediction: an empirical evaluation. Int J Epidemiol. 2011;40:1094-105. [PMID: 21325392]

106. Peters SA, Bakker M, den Ruijter HM, Bots ML. Added value of CAC in risk stratification for cardiovascular events: a systematic review. Eur J Clin Invest. 2012;42:110-6. [PMID: 21644944]

107. Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, et al; International Diagnostic and Prognosis Prediction (IDAPP) Group. Framework for the impact analysis and implementation

of clinical prediction rules (CPRs). BMC Med Inform Decis Mak. 2011;11:62. [PMID: 21999201]

108. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. BMC Med. 2012;10:51. [PMID: 22642691]

109. Campbell MK, Elbourne DR, Altman DG; CONSORT Group. CONSORT statement: extension to cluster randomised trials. BMJ. 2004;328:702-8. [PMID: 15031246]

110. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet. 1974;2:81-4. [PMID: 4136544]

111. Farrell B, Godwin J, Richards S, Warlow C. The United Kingdom transient ischaemic attack (UK-TIA) aspirin trial: final results. J Neurol Neurosurg Psychiatry. 1991;54:1044-54. [PMID: 1783914]

112. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer; 2001.

113. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med. 2010;16: e1000217. [PMID: 20169112]

114. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis: the TRIPOD statement. Ann Intern Med. 2014;162: 55-63.

115. Morise AP, Haddad WJ, Beckner D. Development and validation of a clinical score to estimate the probability of coronary artery disease in men and women presenting with suspected coronary disease. Am J Med. 1997;102:350-6. [PMID: 9217616]

116. Dehing-Oberije C, Yu S, De Ruysscher D, Meersschout S, Van Beek K, Lievens Y, et al. Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys. 2009;74:355-62. [PMID: 19095367]

117. 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:e4181. [PMID: 22723603]

118. Michikawa T, Inoue M, Sawada N, Iwasaki M, Tanaka Y, Shimazu T, et al; Japan Public Health Center-based Prospective Study Group. Development of a prediction model for 10-year risk of hepatocellular carcinoma in middle-aged Japanese: the Japan Public Health Center-based Prospective Study Cohort II. Prev Med. 2012; 55:137-43. [PMID: 22676909]

119. Morise AP, Detrano R, Bobbio M, Diamond GA. Development and validation of a logistic regression-derived algorithm for estimating the incremental probability of coronary artery disease before and after exercise testing. J Am Coll Cardiol. 1992;20:1187-96. [PMID: 1401621]

120. D’Agostino RB Sr, Grundy S, Sullivan LM, Wilson P; CHD Risk Prediction Group. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. JAMA. 2001;286:180-7. [PMID: 11448281]

121. Beck DH, Smith GB, Pappachan JV, Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med. 2003;29: 249-56. [PMID: 12536271]

122. 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;14:40. [PMID: 24645774]

123. Perel P, Prieto-Merino D, Shakur H, Clayton T, Lecky F, Bouamra O, et al. Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. BMJ. 2012; 345:e5166. [PMID: 22896030]

124. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Reardon M, et al. Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. JAMA. 1993; 269:1127-32. [PMID: 8433468]

125. Holland JL, Wilczynski NL, Haynes RB; Hedges Team. Optimal search strategies for identifying sound clinical prediction studies in EMBASE. BMC Med Inform Decis Mak. 2005;5:11. [PMID: 15862125]

126. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J AmMed Inform Assoc. 2001;8:391-7. [PMID: 11418546]

127. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R; Hedges Team. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE. AMIA Annu Symp Proc. 2003:728-32. [PMID: 14728269]

128. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One. 2012;7:e32844. [PMID: 22393453]

129. Keogh C, Wallace E, O’Brien KK, Murphy PJ, Teljeur C, McGrath B, et al. Optimized retrieval of primary care clinical prediction rules from MEDLINE to establish a Web-based register. J Clin Epidemiol. 2011;64:848-60. [PMID: 21411285]

130. Rietveld RP, ter Riet G, Bindels PJ, Sloos JH, van Weert HC. Predicting bacterial cause in infectious conjunctivitis: cohort study on informativeness of combinations of signs and symptoms. BMJ. 2004; 329:206-10. [PMID: 15201195]

131. Poorten VV, Hart A, Vauterin T, Jeunen G, Schoenaers J, Hamoir M, et al. Prognostic index for patients with parotid carcinoma: international external validation in a Belgian-German database. Cancer. 2009;115:540-50. [PMID: 9137571]

132. Moynihan R, Glassock R, Doust J. Chronic kidney disease controversy: how expanding definitions are unnecessarily labelling many people as diseased. BMJ. 2013;347:f4298. [PMID: 23900313]

133. Moynihan R, Henry D, Moons KG. Using evidence to combat overdiagnosis and overtreatment: evaluating treatments, tests, and disease definitions in the time of too much. PLoS Med. 2014;11: e1001655. [PMID: 24983872]

134. Dowling S, Spooner CH, Liang Y, Dryden DM, Friesen C, Klassen TP, et al. Accuracy of Ottawa Ankle Rules to exclude fractures of the ankle and midfoot in children: a meta-analysis. Acad Emerg Med. 2009;16:277-87. [PMID: 19187397]

135. Bachmann LM, Kolb E, Koller MT, Steurer J, ter Riet G. Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid-foot: systematic review. BMJ. 2003;326:417. [PMID: 12595378]

136. Bu¨ ller HR, Ten Cate-Hoek AJ, Hoes AW, Joore MA, Moons KG, Oudega R, et al; AMUSE (Amsterdam Maastricht Utrecht Study on thromboEmbolism) Investigators. Safely ruling out deep venous thrombosis in primary care. Ann Intern Med. 2009;150:229-35. [PMID: 19221374]

137. Sparks AB, Struble CA, Wang ET, Song K, Oliphant A. Noninvasive prenatal detection and selective analysis of cell-free DNA obtained from maternal blood: evaluation for trisomy 21 and trisomy 18. Am J Obstet Gynecol. 2012;206:319.e1-9. [PMID: 22464072]

138. Ankerst DP, Boeck A, Freedland SJ, Thompson IM, Cronin AM, Roobol MJ, et al. Evaluating the PCPT risk calculator in ten international biopsy cohorts: results from the Prostate Biopsy Collaborative Group. World J Urol. 2012;30:181-7. [PMID: 22210512]

139. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008; 336:1475-82. [PMID: 18573856]

140. Conroy RM, Pyo¨ ra¨ la¨ K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al; SCORE Project Group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24:987-1003.

141. Califf RM, Woodlief LH, Harrell FE Jr, Lee KL, White HD, Guerci A, et al. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. Am Heart J. 1997;133:630-9. [PMID: 9200390]

142. McCowan C, Donnan PT, Dewar J, Thompson A, Fahey T. Identifying suspected breast cancer: development and validation of a clinical prediction rule. Br J Gen Pract. 2011;61:e205-14. [PMID: 21619744]

143. Campbell HE, Gray AM, Harris AL, Briggs AH, Taylor MA. Estimation and external validation of a new prognostic model for predicting recurrence-free survival for early breast cancer patients in the UK. Br J Cancer. 2010;103:776-86. [PMID: 20823886]

144. Wilson PW, D’Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97:1837-47. [PMID: 9603539]

145. Kengne AP, Patel A, Marre M, Travert F, Lievre M, Zoungas S, et al; ADVANCE Collaborative Group. Contemporary model for cardiovascular risk prediction in people with type 2 diabetes. Eur J Cardiovasc Prev Rehabil. 2011;18:393-8. [PMID: 21450612]

146. Appelboam A, Reuben AD, Benger JR, Beech F, Dutson J, Haig S, et al. Elbow extension test to rule out elbow fracture: multicentre, prospective validation and observational study of diagnostic accuracy in adults and children. BMJ. 2008;337:a2428. [PMID: 19066257]

147. Puhan MA, Hansel NN, Sobradillo P, Enright P, Lange P, Hickson D, et al; International COPD Cohorts Collaboration Working Group. Large-scale international validation of the ADO index in subjects with COPD: an individual subject data analysis of 10 cohorts. BMJ Open. 2012;2:6. [PMID: 23242246]

148. Knottnerus JA. The Evidence Base of Clinical Diagnosis. London: BMJ Books; 2002.

149. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol. 2003;56: 1118-28. [PMID: 14615003]

150. Grobbee DE, Hoes AW. Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research. London: Jones and Bartlett Publishers; 2009.

151. Sackett DL, Tugwell P, Guyatt GH. Clinical Epidemiology: A Basic Science for Clinical Medicine. 2d ed. Boston: Little, Brown; 1991.

152. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. BMC Med Res Methodol. 2008;8:48. [PMID: 18644127]

153. Knottnerus JA, Dinant GJ. Medicine based evidence, a prerequisite for evidence based medicine. BMJ. 1997;315:1109-10. [PMID: 9374881]

154. Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. BMJ. 2002;324:477-80. [PMID: 11859054]

155. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51:1335-41. [PMID: 15961549]

156. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6. [PMID: 10493205]

157. van Zaane B, Vergouwe Y, Donders AR, Moons KG. Comparison of approaches to estimate confidence intervals of post-test probabilities of diagnostic test results in a nested case-control study. BMC Med Res Methodol. 2012;12:166. [PMID: 23114025]

158. Lumbreras B, Parker LA, Porta M, Polla´n M, Ioannidis JP, Herna´ndez-Aguado I. Overinterpretation of clinical applicability in molecular diagnostic research. Clin Chem. 2009;55:786-94. [PMID: 19233907]

159. Tzoulaki I, Siontis KC, Ioannidis JP. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. BMJ. 2011;343: d6829. [PMID: 22065657]

160. Greving JP, Wermer MJ, Brown RD Jr, Morita A, Juvela S, Yonekura M, et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. Lancet Neurol. 2014;13:59-66. [PMID: 24290159]

161. Collins GS, Altman DG. Predicting the adverse risk of statin treatment: an independent and external validation of Qstatin risk scores in the UK. Heart. 2012;98:1091-7. [PMID: 22689714]

162. Glickman SW, Shofer FS, Wu MC, Scholer MJ, Ndubuizu A, Peterson ED, et al. Development and validation of a prioritization rule for obtaining an immediate 12-lead electrocardiogram in the emergency department to identify ST-elevation myocardial infarction. Am Heart J. 2012;163:372-82. [PMID: 22424007]

163. Debray TP, Koffijberg H, Lu D, Vergouwe Y, Steyerberg EW, Moons KG. Incorporating published univariable associations in diagnostic and prognostic modeling. BMC Med Res Methodol. 2012;12: 121. [PMID: 22883206]

164. Debray TP, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. Stat Med. 2012; 31:2697-712. [PMID: 22733546]

165. Debray TP, Moons KG, Abo-Zaid GM, Koffijberg H, Riley RD. Individual participant data meta-analysis for a binary outcome: onestage or two-stage? PLoS One. 2013;8:e60650. [PMID: 23585842]

166. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32:3158-80. [PMID: 23307585]

167. Bouwmeester W, Twisk JW, Kappen TH, van Klei WA, Moons KG, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC Med Res Methodol. 2013;13:19. [PMID: 23414436]

168. Bouwmeester W, Moons KG, Happen TH, van Klei WA, Twisk JW, Eijkemans MJ, et al. Internal validation of risk models in clustered data: a comparison of bootstrap schemes. Am J Epidemiol. 2013;177:1209-17. [PMID: 23660796]

169. Rosner B, Qiu W, Lee ML. Assessing discrimination of risk prediction rules in a clustered data setting. Lifetime Data Anal. 2013;19: 242-56. [PMID: 23263872]

170. van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. BMC Med Res Methodol. 2014;14:5. [PMID: 24423445]

171. van Klaveren D, Steyerberg EW, Vergouwe Y. Interpretation of concordance measures for clustered data. Stat Med. 2014;33:714-6. [PMID: 24425541]

172. Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L. Derivation and assessment of risk prediction models using casecohort data. BMC Med Res Methodol. 2013;13:113. [PMID: 24034146]

173. Ganna A, Reilly M, de Faire U, Pedersen N, Magnusson P, Ingelsson E. Risk prediction measures for case-cohort and nested case-control designs: an application to cardiovascular disease. Am J Epidemiol. 2012;175:715-24. [PMID: 22396388]

174. Kulathinal S, Karvanen J, Saarela O, Kuulasmaa K. Case-cohort design in practice—experiences from the MORGAM Project. Epidemiol Perspect Innov. 2007;4:15. [PMID: 18053196]

175. Kengne AP, Beulens JW, Peelen LM, Moons KG, van der Schouw YT, Schulze MB, et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models. Lancet Diabetes Endocrinol. 2014;2:19-29. [PMID: 24622666]

176. Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, et al. Risk prediction models for mortality in ambulatory heart failure patients: a systematic review. Circ Heart Fail. 2013;6: 881-9. [PMID: 23888045]

177. Arkenau HT, Barriuso J, Olmos D, Ang JE, de Bono J, Judson I, et al. Prospective validation of a prognostic score to improve patient selection for oncology phase I trials. J Clin Oncol. 2009;27:2692-6. [PMID: 19332724]

178. Ronga A, Vaucher P, Haasenritter J, Donner-Banzhoff N, Bo¨ sner S, Verdon F, et al. Development and validation of a clinical prediction rule for chest wall syndrome in primary care. BMC Fam Pract. 2012;13:74. [PMID: 22866824]

179. Martinez JA, Belastegui A, Basabe I, Goicoechea X, Aguirre C, Lizeaga N, et al. Derivation and validation of a clinical prediction rule for delirium in patients admitted to a medical ward: an observational study. BMJ Open. 2012;2:e001599. [PMID: 22983876]

180. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, et al. Risk prediction in patients with heart failure: a systematic review and analysis. JACC Heart Fail. 2014;2:440-6. [PMID: 25194291]

181. Ebell MH, Afonson AM, Gonzales R, Stein J, Genton B, Senn N. Development and validation of a clinical decision rule for the diagnosis of influenza. J Am Board Fam Med. 2012;25:55-62. [PMID: 22218625]

182. Counsell C, Dennis M. Systematic review of prognostic models in patients with acute stroke. Cerebrovasc Dis. 2001;12:159-70. [PMID: 11641579]

183. Knottnerus JA. Between iatrotropic stimulus and interiatric referral: the domain of primary care research. J Clin Epidemiol. 2002; 55:1201-6. [PMID: 12547450]

184. Moreno R, Apolone G. Impact of different customization strategies in the performance of a general severity score. Crit Care Med. 1997;25:2001-8. [PMID: 9403750]

185. Tu JV, Austin PC, Walld R, Roos L, Agras J, McDonald KM. Development and validation of the Ontario acute myocardial infarction mortality prediction rules. J Am Coll Cardiol. 2001;37:992-7. [PMID: 11263626]

186. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172:971-80. [PMID: 20807737]

187. Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KG. Adaptation of clinical prediction models for application in local settings. Med Decis Making. 2012;32:E1-10. [PMID: 22427369]

188. Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. Ann Intern Med. 2005;143:100-7. [PMID: 16027451]

189. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol. 1992;45: 1143-54. [PMID: 1474411]

190. Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and diseases. Med Decis Making. 1987;7:139-48. [PMID: 3613914]

191. Eberhart LH, Morin AM, Guber D, Kretz FJ, Scha¨ uffelen A, Treiber H, et al. Applicability of risk scores for postoperative nausea and vomiting in adults to paediatric patients. Br J Anaesth. 2004;93: 386-92. [PMID: 15247114]

192. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2014 Aug 29. [Epub ahead of print]. [PMID: 25179855]

193. Klemke CD, Mansmann U, Poenitz N, Dippel E, Goerdt S. Prognostic factors and prediction of prognosis by the CTCL Severity Index in mycosis fungoides and Se´ zary syndrome. Br J Dermatol. 2005;153:118-24. [PMID: 16029336]

194. Tay SY, Thoo FL, Sitoh YY, Seow E, Wong HP. The Ottawa Ankle Rules in Asia: validating a clinical decision rule for requesting X-rays in twisting ankle and foot injuries. J Emerg Med. 1999;17:945-7. [PMID: 10595876]

195. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087-91. [PMID: 16980149]

196. Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ. 2012;184:1265-9. [PMID: 22371511]

197. Janssen KJ, Donders AR, Harrell FE, Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol. 2010;63:721-7. [PMID: 20338724]

198. Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem. 2009;55:994-1001. [PMID: 19282357]

199. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59:1092-101. [PMID: 16980150]

200. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. [PMID: 19564179]

201. Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63:205-14. [PMID: 19596181]

202. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al; PROGRESS Group. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346: 35595. [PMID: 23386360]

203. Liew SM, Doust J, Glasziou P. Cardiovascular risk scores do not account for the effect of treatment: a review. Heart. 2011;97:689-97. [PMID: 21474616]

204. Simon R, Altman D, G. Statistical aspects of prognostic factor studies in oncology. Br J Cancer. 1994;69:979-85. [PMID: 8198989]

205. Landefeld CS, Goldman L. Major bleeding in outpatients treated with warfarin: incidence and prediction by factors known at the start of outpatient therapy. Am J Med. 1989;87:144-52. [PMID: 2787958]

206. Schuit E, Groenwold RH, Harrell FE Jr, de Kort WL, Kwee A, Mol BW, et al. Unexpected predictor-outcome associations in clinical prediction research: causes and solutions. CMAJ. 2013;185:E499-505. [PMID: 23339155]

207. Wong J, Taljaard M, Forster AJ, Escobar GJ, van Walraven C. Addition of time-dependent covariates to a survival model significantly improved predictions for daily risk of hospital death. J Eval Clin Pract. 2013;19:351-7. [PMID: 22409151]

208. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation

of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA. 2007;297: 611-9. [PMID: 17299196]

209. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009;62:797-806. [PMID: 19447581]

210. Massing MW, Simpson RJ Jr, Rautaharju PM, Schreiner PJ, Crow R, Heiss G. Usefulness of ventricular premature complexes to predict coronary heart disease events and mortality (from the Atherosclerosis Risk In Communities cohort). Am J Cardiol. 2006;98:1609-12. [PMID: 17145219]

211. Craig JC, Williams GJ, Jones M, Codarini M, Macaskill P, Hayen A, et al. The accuracy of clinical symptoms and signs for the diagnosis of serious bacterial infection in young febrile children: prospective cohort study of 15 781 febrile illnesses. BMJ. 2010;340:c1594. [PMID: 20406860]

212. Todenhofer T, Renninger M, Schwentner C, Stenzl A, Gakis G. A new prognostic model for cancer-specific survival after radical cystectomy including pretreatment thrombocytosis and standard pathological risk factors. BJU Int. 2012;110(11 Pt B):E533-40. [PMID: 22578156]

213. Boggs DA, Rosenberg L, Pencina MJ, Adams-Campbell LL, Palmer JR. Validation of a breast cancer risk prediction model developed for Black women. J Natl Cancer Inst. 2013;105:361-7. [PMID: 23411594]

214. Knottnerus JA, Buntinx F. The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research. Hoboken, NJ: Wiley-Blackwell; 2009.

215. Naaktgeboren CA, de Groot JA, van Smeden M, Moons KG, Reitsma JB. Evaluating diagnostic accuracy in the face of multiple reference standards. Ann Intern Med. 2013;159:195-202. [PMID: 23922065]

216. Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med. 2013;10:e1001531. [PMID: 24143138]

217. Naaktgeboren CA, Bertens LC, van Smeden M, Groot JA, Moons KG, Reitsma JB. Value of composite reference standards in diagnostic research. BMJ. 2013;347:f5605. [PMID: 24162938]

218. de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. 2011;343:d4770. [PMID: 21810869]

219. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Brophy J, Joseph L, et al. Adjusting for partial verification or workup bias in meta-analyses of diagnostic accuracy studies. Am J Epidemiol. 2012;175:847-53. [PMID: 22422923]

220. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174:469-76. [PMID: 16477057]

221. Rouzier R, Pusztai L, Delaloge S, Gonzalez-Angulo AM, Andre F, Hess KR, et al. Nomograms to predict pathologic complete response and metastasis-free survival after preoperative chemotherapy for breast cancer. J Clin Oncol. 2005;23:8331-9. [PMID: 16293864]

222. Elliott J, Beringer T, Kee F, Marsh D, Willis C, Stevenson M. Predicting survival after treatment for fracture of the proximal femur and the effect of delays to surgery. J Clin Epidemiol. 2003;56:788-95. [PMID: 12954472]

223. Adams LA, Bulsara M, Rossi E, DeBoer B, Speers D, George J, et al. Hepascore: an accurate validated predictor of liver fibrosis in chronic hepatitis C infection. Clin Chem. 2005;51:1867-73. [PMID: 16055434]

224. Hess EP, Brison RJ, Perry JJ, Calder LA, Thiruganasambandamoorthy V, Agarwal D, et al. Development of a clinical prediction rule for 30-day cardiac events in emergency department patients with chest pain and possible acute coronary syndrome. Ann Emerg Med. 2012;59:115-25. [PMID: 21885156]

225. Moons KG, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic studies? J Clin Epidemiol. 2002;55:633-6. [PMID: 12160909]

226. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007;iii, ix-51. [PMID: 18021577]

227. Kaijser J, Sayasneh A, Van Hoorde K, Ghaem-Maghami S, Bourne T, Timmerman D, et al. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. Hum Reprod Update. 2014;20:449-52. [PMID: 24327552]

228. Kaul V, Friedenberg FK, Braitman LE, Anis U, Zaeri N, Fazili J, et al. Development and validation of a model to diagnose cirrhosis in patients with hepatitis C. Am J Gastroenterol. 2002;97:2623-8. [PMID: 12385450]

229. Halbesma N, Jansen DF, Heymans MW, Stolk RP, de Jong PE, Gansevoort RT ; PREVEND Study Group. Development and validation of a general population renal risk score. Clin J Am Soc Nephrol. 2011;6:1731-8. [PMID: 21734089]

230. Beyersmann J, Wolkewitz M, Schumacher M. The impact of time-dependent bias in proportional hazards modelling. Stat Med. 2008;27:6439-54. [PMID: 18837068]

231. van Walraven C, Davis D, Forster AJ, Wells GA. Timedependent bias was common in survival analyses published in leading clinical journals. J Clin Epidemiol. 2004;57:672-82. [PMID: 15358395]

232. Rochon J. Issues in adjusting for covariates arising postrandomization in clinical trials. Drug Inf J. 1999;33:1219-28.

233. D’Agostino RB. Beyond baseline data: the use of time-varying covariates. J Hypertens. 2008;26:639-40. [PMID: 18327070]

234. Scheike TH. Time-varying effects in survival analysis. In: Rao CR, ed. Advances in Survival Analysis. Amsterdam: Elsevier; 2004:61-8.

235. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49:907-16. [PMID: 8699212]

236. Rutten FH, Vonken EJ, Cramer MJ, Moons KG, Velthuis BB, Prakken NH, et al. Cardiovascular magnetic resonance imaging to identify left-sided chronic heart failure in stable patients with chronic obstructive pulmonary disease. Am Heart J. 2008;156:506-12. [PMID: 18760133]

237. Hess EP, Perry JJ, Calder LA, Thiruganasambandamoorthy V, Body R, Jaffe A, et al. Prospective validation of a modified thrombolysis in myocardial infarction risk score in emergency department patients with chest pain and possible acute coronary syndrome. Acad Emerg Med. 2010;17:368-75. [PMID: 20370775]

238. Begg CB. Bias in the assessment of diagnostic tests. Stat Med. 1987;6:411-23. [PMID: 3114858]

239. Elmore JG, Wells CK, Howard DH, Feinstein AR. The impact of clinical history on mammographic interpretations. JAMA. 1997;277:49-52. [PMID: 8980210]

240. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. JAMA. 2004;292:1602-9. [PMID: 15467063]

241. Loewen P, Dahir K. Risk of bleeding with oral anticoagulants: an updated systematic review and performance analysis of clinical prediction rules. Ann Hematol. 2011;90:1191-200. [PMID: 21670974]

242. Sheth T, Butler C, Chow B, Chan MT, Mitha A, Nagele P, et al; CTA VISION Investigators. The coronary CT angiography vision protocol: a prospective observational imaging cohort study in patients undergoing non-cardiac surgery. BMJ Open. 2012;2:e001474. [PMID: 22855630]

243. Hippisley-Cox J, Coupland C. Identifying patients with suspected pancreatic cancer in primary care: derivation and validation of an algorithm. Br J Gen Pract. 2012;62:e38-e45. [PMID: 22520674]

244. Holmes JF, Mao A, Awasthi S, McGahan JP, Wisner DH, Kuppermann N. Validation of a prediction rule for the identification of children with intra-abdominal injuries after blunt torso trauma. Ann Emerg Med. 2009;54:528-33. [PMID: 19250706]

245. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48:1503-12. [PMID: 8543964]

246. 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;49:1373-9. [PMID: 8970487]

247. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007;165:710-8. [PMID: 17182981]

248. Feinstein AR. Multivariable Analysis. New Haven, CT: Yale University Press; 1996.

249. Schumacher M, Holla¨nder N, Schwarzer G, Binder H, Sauerbrei W. Prognostic factor studies. In: Crowley J, Hoering A, eds. Handbook of Statistics in Clinical Oncology. 3rd ed. London: Chapman and Hall/CRC; 2012:415-70.

250. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64:993-1000. [PMID: 21411281]

251. Jinks RC. Sample size for multivariable prognostic models. PhD thesis. University College London; 2012.

252. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski

N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010; 21:128-38. [PMID: 20010215]

253. Steyerberg EW, Calster BV, Pencina MJ. Performance measures for prediction models and markers: evaluation of predictions and classifications. Rev Esp Cardiol (Engl Ed). 2011;64:788-94. [PMID: 24775954]

254. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58:475-83. [PMID: 15845334]

255. Audige´ L, Bhandari M, Kellam J. How reliable are reliability studies of fracture classifications? A systematic review of their methodologies. Acta Orthop Scand. 2004;75:184-94. [PMID: 15180234]

256. Genders TS, Steyerberg EW, Hunink MG, Nieman K, Galema TW, Mollet NR, et al. Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. BMJ. 2012;344:e3485. [PMID: 22692650]

257. Thompson DO, Hurtado TR, Liao MM, Byyny RL, Gravitz C, Haukoos JS. Validation of the Simplified Motor Score in the out-ofhospital setting for the prediction of outcomes after traumatic brain injury. Ann Emerg Med. 2011;58:417-25. [PMID: 21803448]

258. Ambler G, Omar RZ, Royston P, Kinsman R, Keogh BE, Taylor KM. Generic, simple risk stratification model for heart valve surgery. Circulation. 2005;112:224-31. [PMID: 15998680]

259. Mackinnon A. The use and reporting of multiple imputation in medical research—a review. J Intern Med. 2010;268:586-93. [PMID: 20831627]

260. Hussain A, Dunn KW. Predicting length of stay in thermal burns: a systematic review of prognostic factors. Burns. 2013;39:1331-40. [PMID: 23768707]

261. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. JAMA. 2011;305:1553-9. [PMID: 21482743]

262. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med. 2008;5:e165. [PMID: 18684008]

263. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung cancer risk prediction: Prostate, Lung, Colorectal And Ovarian Cancer Screening Trial models and validation. J Natl Cancer Inst. 2011;103:1058-68. [PMID: 21606442]

264. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using “optimal” cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86:829-35. [PMID: 8182763]

265. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25:127-41. [PMID: 16217841]

266. Royston P, Sauerbrei W. Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables. Chichester: John Wiley;2008.

267. Veerbeek JM, Kwakkel G, van Wegen EE, Ket JC, Heymans MW. Early prediction of outcome of activities of daily living after stroke: a systematic review. Stroke. 2011;42:1482-8. [PMID: 21474812]

268. Lubetzky-Vilnai A, Ciol M, McCoy SW. Statistical analysis of clinical prediction rules for rehabilitation interventions: current state of the literature. Arch Phys Med Rehabil. 2014;95:188-96. [PMID: 24036159]

269. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35:1925-31. [PMID: 24898551]

270. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008;19:640-8. [PMID: 18633328]

271. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials. 2010;11:9. [PMID: 20113465]

272. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: Wiley; 2000.

273. Vittinghoff E. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. New York: Springer;2005.

274. Hosmer DW, Lemeshow S, May S. Applied Survival Analysis: Regression Modelling of Time-To-Event Data. Hoboken, NJ: Wiley-Interscience; 2008.

275. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2001.

276. Kuhn M, Johnson K. Applied Predictive Modelling. New York: Springer; 2013.

277. Andersen PK, Skovgaard LT. Regression With Linear Predictors. New York: Springer; 2010.

278. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007;335:136. [PMID: 17615182]

279. Moreno L, Krishnan JA, Duran P, Ferrero F. Development and validation of a clinical prediction rule to distinguish bacterial from viral pneumonia in children. Pediatr Pulmonol. 2006;41:331-7. [PMID: 16493666]

280. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. Am Heart J. 1991;121:293-8. [PMID: 1985385]

281. Royston P, Parmar MK. Flexible parametric proportionalhazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21:2175-97. [PMID: 12210632]

282. Hans D, Durosier C, Kanis JA, Johansson H, Schott-Pethelaz AM, Krieg MA. Assessment of the 10-year probability of osteoporotic hip fracture combining clinical risk factors and heel bone ultrasound: the EPISEM prospective cohort of 12,958 elderly women. J Bone Miner Res. 2008;23:1045-51. [PMID: 18302507]

283. Bohensky MA, Jolley D, Pilcher DV, Sundararajan V, Evans S, Brand CA. Prognostic models based on administrative data alone inadequately predict the survival outcomes for critically ill patients at 180 days post-hospital discharge. J Crit Care. 2012;27:422.e11-21. [PMID: 22591572]

284. Barrett TW, Martin AR, Storrow AB, Jenkins CA, Harrell FE, Russ S, et al. A clinical prediction model to estimate risk for 30-day adverse events in emergency department patients with symptomatic atrial fibrillation. Ann Emerg Med. 2011;57:1-12. [PMID: 20728962]

285. Krijnen P, van Jaarsveld BC, Steyerberg EW, Man in ‘t Veld AJ, Schalekamp MA, Habbema JD. A clinical prediction rule for renal artery stenosis. Ann Intern Med. 1998;129:705-11. [PMID: 9841602]

286. Smits M, Dippel DW, Steyerberg EW, de Haan GG, Dekker HM, Vos PE, et al. Predicting intracranial traumatic findings on computed tomography in patients with minor head injury: the CHIP prediction rule. Ann Intern Med. 2007;146:397-405. [PMID: 17371884]

287. Moons KG, Donders AR, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol. 2004;57:1262-70. [PMID: 15617952]

288. Mantel N. Why stepdown procedures in variable selection? Technometrics. 1970;12:621-5.

289. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003;56:826-32. [PMID: 14505766]

290. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med. 2004;23:2567-86. [PMID: 15287085]

291. van Houwelingen HC, Sauerbrei W. Cross-validation, shrinkage and variable selection in linear regression revisited. Open J Statist. 2013;3:79-102.

292. Sauerbrei W, Boulesteix AL, Binder H. Stability investigations of multivariable regression models derived from low- and highdimensional data. J Biopharm Stat. 2011;21:1206-31. [PMID: 22023687]

293. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med. 1984;3:143-52. [PMID: 6463451]

294. van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med. 1990;9:1303-25. [PMID: 2277880]

295. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005;21: 3301-7. [PMID: 15905277]

296. Chatfield C. Model uncertainty, data mining and statistical inference. J R Stat Soc A. 1995;158:419-66.

297. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007;26:5512-28. [PMID: 18058845]

298. Heymans MW, van Buuren S, Knol DL, van Mechelen W, de Vet HC. Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Med Res Meth. 2007;7:33. [PMID: 17629912]

299. Castaldi PJ, Dahabreh IJ, Ioannidis JP. An empirical assessment of validation practices for molecular classifiers. Brief Bioinform. 2011; 12:189-202. [PMID: 21300697]

300. Varma S, Simon R. Bias in error estimation when using crossvalidation for model selection. BMC Bioinformatics. 2006;7:91. [PMID: 16504092]

301. Vach K, Sauerbrei W, Schumacher M. Variable selection and shrinkage: comparison of some approaches. Stat Neerl. 2001;55:53-75.

302. Lin IF, Chang WP, Liao YN. Shrinkage methods enhanced the accuracy of parameter estimation using Cox models with small number of events. J Clin Epidemiol. 2013;66:743-51. [PMID: 23566374]

303. Ambler G, Seaman S, Omar RZ. An evaluation of penalized survival methods for developing prognostic models with rare events. Stat Med. 2012;31:1150-61. [PMID: 21997569]

304. Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. Prognostic indices for older adults: a systematic review. JAMA. 2012; 307:182-92. [PMID: 22235089]

305. Spelt L, Andersson B, Nilsson J, Andersson R. Prognostic models for outcome following liver resection for colorectal cancer metastases: a systematic review. Eur J Surg Oncol. 2012;38:16-24. [PMID: 22079259]

306. Nam RK, Kattan MW, Chin JL, Trachtenberg J, Singal R, Rendon R, et al. Prospective multi-institutional study evaluating the performance of prostate cancer risk calculators. J Clin Oncol. 2011;29: 2959-64. [PMID: 21690464]

307. Meffert PJ, Baumeister SE, Lerch MM, Mayerle J, Kratzer W, Vo¨ lzke H. Development, external validation, and comparative assessment of a new diagnostic score for hepatic steatosis. Am J Gastroenterol. 2014;109:1404-14. [PMID: 24957156]

308. Collins GS, Altman DG. Identifying patients with undetected colorectal cancer: an independent validation of QCancer (Colorectal). Br J Cancer. 2012;107:260-5. [PMID: 22699822]

309. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13: 33. [PMID: 23496923]

310. Iasonos A, Schrag D, Raj GV, Panageas KS. How to build and interpret a nomogram for cancer prognosis. J Clin Oncol. 2008;26: 1364-70. [PMID: 18323559]

311. Zivanovic O, Jacks LM, Iasonos A, Leitao MM Jr, Soslow RA, Veras E, et al. A nomogram to predict postresection 5-year overall survival for patients with uterine leiomyosarcoma. Cancer. 2012;118:660-9. [PMID: 21751199]

312. Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, et al. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. Osteoporos Int. 2007;18:1033-46. [PMID: 17323110]

313. Papaioannou A, Morin S, Cheung AM, Atkinson S, Brown JP, Feldman S, et al; Scientific Advisory Council of Osteoporosis Canada. 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. CMAJ. 2010;182:1864-73. [PMID: 20940232]

314. Collins GS, Michae¨ lsson K. Fracture risk assessment: state of the art, methodologically unsound, or poorly reported? Curr Osteoporos Rep. 2012;10:199-207. [PMID: 22688862]

315. Collins GS, Mallett S, Altman DG. Predicting risk of osteoporotic and hip fracture in the United Kingdom: prospective independent and external validation of QFractureScores. BMJ. 2011;342:d3651. [PMID: 21697214]

316. Ja¨ rvinen TL, Jokihaara J, Guy P, Alonso-Coello P, Collins GS, Michae¨ lsson K, et al. Conflicts at the heart of the FRAX tool. CMAJ. 2014;186:165-7. [PMID: 24366895]

317. Balman˜ a J, Stockwell DH, Steyerberg EW, Stoffel EM, Deffenbaugh AM, Reid JE, et al. Prediction of MLH1 and MSH2 mutations in Lynch syndrome. JAMA. 2006;296:1469-78. [PMID: 17003395]

318. Bruins Slot MH, Rutten FH, van der Heijden GJ, Geersing GJ, Glatz JF, Hoes AW. Diagnosing acute coronary syndrome in primary care: comparison of the physicians' risk estimation and a clinical decision rule. Fam Pract. 2011;28:323-8. [PMID: 21239470]

319. Suarthana E, Vergouwe Y, Moons KG, de Monchy J, Grobbee D, Heederik D, et al. A diagnostic model for the detection of sensitization to wheat allergens was developed and validated in bakery workers. J Clin Epidemiol. 2010;63:1011-9. [PMID: 20189762]

320. Uno H, Cai T, Pencina MJ, D’Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30:1105-17. [PMID: 21484848]

321. Akazawa K. Measures of explained variation for a regression model used in survival analysis. J Med Syst. 1997;21:229-38. [PMID: 9442437]

322. Choodari-Oskooei B, Royston P, Parmar MK. A simulation study of predictive ability measures in a survival model I: explained variation measures. Stat Med. 2012;31:2627-43. [PMID: 21520455]

323. Heller G. A measure of explained risk in the proportional hazards model. Biostatistics. 2012;13:315-25. [PMID: 22190711]

324. Korn EL, Simon R. Measures of explained variation for survival data. Stat Med. 1990;9:487-503. [PMID: 2349402]

325. Mittlbo¨ ck M, Schemper M. Explained variation for logistic regression. Stat Med. 1996;15:1987-97. [PMID: 8896134]

326. Royston P. Explained variation for survival models. Stata Journal. 2006;6:83-96

327. Schemper M. Predictive accuracy and explained variation. Stat Med. 2003;22:2299-308. [PMID: 12854094]

328. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. Biometrics. 2000;56:249-55. [PMID: 10783803]

329. Schemper M, Stare J. Explained variation in survival analysis. Stat Med. 1996;15:1999-2012. [PMID: 8896135]

330. Gerds T, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biom J. 2006;6:1029-40. [PMID: 17240660]

331. Rufibach K. Use of Brier score to assess binary predictions. J Clin Epidemiol. 2010;63:938-9. [PMID: 20189763]

332. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biom J. 2008;50:457-79. [PMID: 18663757]

333. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med. 2004;23:723-48. [PMID: 14981672]

334. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837-45. [PMID: 3203132]

335. Demler OV, Pencina MJ, D’Agostino RB Sr. Misuse of DeLong test to compare AUCs for nested models. Stat Med. 2012;31:2577-87. [PMID: 22415937]

336. Moonesinghe SR, Mythen MG, Das P, Rowan KM, Grocott MP. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. Anesthesiology. 2013;119:959-81. [PMID: 24195875]

337. Wallace E, Stuart E, Vaughan N, Bennett K, Fahey T, Smith SM. Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. Med Care. 2014;52: 751-65. [PMID: 25023919]

338. Widera C, Pencina MJ, Bobadilla M, Reimann I, Guba-Quint A, Marquardt I, et al. Incremental prognostic value of biomarkers beyond the GRACE (Global Registry of Acute Coronary Events) score and high-sensitivity cardiac troponin T in non-ST-elevation acute coronary syndrome. Clin Chem. 2013;59:1497-505. [PMID: 23818444]

339. Pencina MJ, D’Agostino RB Sr, D’Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27:157-72. [PMID: 17569110]

340. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115:928-35. [PMID:17309939]

341. Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, et al; American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation. 2009;119:2408-16. [PMID: 19364974]

342. Cook NR. Assessing the incremental role of novel and emerging risk factors. Curr Cardiovasc Risk Rep. 2010;4:112-9. [PMID: 20640227]

343. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol. 2011;11:13. [PMID: 21276237]

344. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med. 2009;150:795-802. [PMID: 19487714]

345. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. Biom J. 2011;53:237-58. [PMID: 21294152]

346. Cook NR. Clinically relevant measures of fit? A note of caution. Am J Epidemiol. 2012;176:488-91. [PMID: 22875759]

347. Pencina MJ, D’Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. Am J Epidemiol. 2012;176:473-81. [PMID: 22875755]

348. Pencina MJ, D’Agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. Clin Chem Lab Med. 2010;48:1703-11. [PMID: 20716010]

349. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of markers and risk prediction models: overview of relationships between NRI and decision-analytic measures. Med Decis Making. 2013;33:490-501. [PMID: 23313931]

350. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. Stat Med. 2014;33:3405-14. [PMID: 23553436]

351. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. Am J Epidemiol. 2011;173:1327-35. [PMID: 21555714]

352. Mihaescu R, van Zitteren M, van Hoek M, Sijbrands EJ, Uitterlinden AG, Witteman JC, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. Am J Epidemiol. 2010; 172:353-61. [PMID: 20562194]

353. Mu¨ hlenbruch K, Heraclides A, Steyerberg EW, Joost HG, Boeing H, Schulze MB. Assessing improvement in disease prediction using net reclassification improvement: impact of risk cut-offs and number of risk categories. Eur J Epidemiol. 2013;28:25-33. [PMID: 23179629]

354. Pepe M, Fang J, Feng Z, Gerds T, Hilden J. The Net Reclassification Index (NRI): a Misleading Measure of Prediction Improvement with Miscalibrated or Overfit Models. UWBiostatistics Working Paper Series. Working Paper 392. Madison, WI: University of Wisconsin; 2013.

355. Vickers AJ, Pepe M. Does the net reclassification improvement help us evaluate models and markers? Ann Intern Med. 2014;160: 136-7. [PMID: 24592500]

356. Hilden J. Commentary: On NRI, IDI, and “good-looking” statistics with nothing underneath. Epidemiology. 2014;25:265-7. [PMID: 24487208]

357. Leening MJ, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med. 2014;160:122-31. [PMID: 24592497]

358. Al-Radi OO, Harrell FE Jr, Caldarone CA, McCrindle BW, Jacobs JP, Williams MG, et al. Case complexity scores in congenital heart surgery: a comparative study of the Aristotle Basic Complexity score and the Risk Adjustment in Congenital Heart Surgery (RACHS-1) system. J Thorac Cardiovasc Surg. 2007;133:865-75. [PMID: 17382616]

359. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med. 2012;157:294-5. [PMID: 22910942]

360. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making. 2014 Aug 25. [Epub ahead of print]. [PMID: 25155798]

361. Vickers AJ. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. Am Stat. 2008;62: 314-20. [PMID: 19132141]

362. Vickers AJ, Cronin AM, Kattan MW, Gonen M, Scardino PT, Milowsky MI, et al; International Bladder Cancer Nomogram Consortium. Clinical benefits of a multivariate prediction model for bladder cancer: a decision analytic approach. Cancer. 2009;115:5460-9. [PMID: 19823979]

363. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74. [PMID: 17099194]

364. Baker SG. Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst. 2009;101:1538-42. [PMID: 19843888]

365. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc. 2009; 172:729-48. [PMID: 20069131]

366. Baker SG, Kramer BS. Evaluating a new marker for risk prediction: decision analysis to the rescue. Discov Med. 2012;14:181-8. [PMID: 23021372]

367. Moons KG, de Groot JA, Linnet K, Reitsma JB, Bossuyt PM. Quantifying the added value of a diagnostic test or marker. Clin Chem. 2012;58:1408-17. [PMID: 22952348]

368. Held U, Bove´ DS, Steurer J, Held L. Validating and updating a risk model for pneumonia—a case study. BMC Med Res Methodol. 2012;12:99. [PMID: 22817850]

369. Cindolo L, Chiodini P, Gallo C, Ficarra V, Schips L, Tostain J, et al. Validation by calibration of the UCLA integrated staging system prognostic model for nonmetastatic renal cell carcinoma after nephrectomy. Cancer. 2008;113:65-71. [PMID: 18473356]

370. Baart AM, Atsma F, McSweeney EN, Moons KG, Vergouwe Y, de Kort WL. External validation and updating of a Dutch prediction model for low hemoglobin deferral in Irish whole blood donors. Transfusion. 2014;54(3 Pt 2):762-9. [PMID: 23607909]

371. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009;374:86-9. [PMID: 19525005]

372. Janssen KJ, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KG. A simple method to adjust clinical prediction models to local circumstances. Can J Anaesth. 2009;56:194-201. [PMID: 19247740]

373. van Houwelingen HC. Validation. calibration, revision and combination of prognostic survival models. Stat Med. 2000;19:3401-15. [PMID: 11122504]

374. Manola J, Royston P, Elson P, McCormack JB, Mazumdar M, Ne´ grier S, et al; International Kidney Cancer Working Group. Prognostic model for survival in patients with metastatic renal cell carcinoma: results from the International Kidney Cancer Working Group. Clin Cancer Res. 2011;17:5443-50. [PMID: 21828239]

375. Krupp NL, Weinstein G, Chalian A, Berlin JA, Wolf P, Weber RS. Validation of a transfusion prediction model in head and neck cancer surgery. Arch Otolaryngol Head Neck Surg. 2003;129:1297-302. [PMID: 14676155]

376. Morra E, Cesana C, Klersy C, Barbarano L, Varettoni M, Cavanna L, et al. Clinical characteristics and factors predicting evolution of asymptomatic IgM monoclonal gammopathies and IgMrelated disorders. Leukemia. 2004;18:1512-7. [PMID: 15322559]

377. Kelder JC, Cramer MJ, van Wijngaarden J, van Tooren R, Mosterd A, Moons KG, et al. The diagnostic value of physical examination and additional testing in primary care patients with suspected heart failure. Circulation. 2011;124:2865-73. [PMID: 22104551]

378. Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al. A prognostic index in primary breast cancer. Br J Cancer. 1982;45:361-6. [PMID: 7073932]

379. Tang EW, Wong CK, Herbison P. Global Registry of Acute Coronary Events (GRACE) hospital discharge risk score accurately predicts long-term mortality post acute coronary syndrome. Am Heart J. 2007;153:29-35. [PMID: 17174633]

380. Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA, et al. Development and validation of a patient selfassessment score for diabetes risk. Ann Intern Med. 2009;151:775-83. [PMID: 19949143]

381. Chen L, Magliano DJ, Balkau B, Colagiuri S, Zimmet PZ, Tonkin AM, et al. AUSDRISK: an Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust. 2010;192:197-202. [PMID: 20170456]

382. Starmans R, Muris JW, Fijten GH, Schouten HJ, Pop P, Knottnerus JA. The diagnostic value of scoring models for organic and nonorganic gastrointestinal disease, including the irritable-bowel syndrome. Med Decis Making. 1994;14:208-16. [PMID: 7934707]

383. Tzoulaki I, Seretis A, Ntzani EE, Ioannidis JP. Mapping the expanded often inappropriate use of the Framingham Risk Score in the medical literature. J Clin Epidemiol. 2014;67:571-7. [PMID: 24513280]

384. Harrison DA, Rowan KM. Outcome prediction in critical care: the ICNARC model. Curr Opin Crit Care. 2008;14:506-12. [PMID: 18787441]

385. Kanaya AM, Wassel Fyr CL, de Rekeneire N, Schwartz AV, Goodpaster BH, Newman AB, et al. Predicting the development of diabetes in older adults: the derivation and validation of a prediction rule. Diabetes Care. 2005;28:404-8. [PMID: 15677800]

386. Stephens JW, Ambler G, Vallance P, Betterridge DJ, Humphries SE, Hurel SJ. Cardiovascular risk and diabetes. Are the methods of risk prediction satisfactory? Eur J Cardiovasc Prev Rehabil. 2004;11:521-8. [PMID: 15580065]

387. Cogswell R, Kobashigawa E, McGlothlin D, Shaw R, De Marco T. Validation of the Registry to Evaluate Early and Long-Term Pulmonary Arterial Hypertension Disease Management (REVEAL) pulmonary hypertension prediction model in a unique population and utility in the prediction of long-term survival. J Heart Lung Transplant. 2012;31:1165-70. [PMID: 23062726]

388. Eagle KA, Lim MJ, Dabbous OH, Pieper KS, Goldberg RJ, Van de Werf F, et al; GRACE Investigators. A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. JAMA. 2004;291:2727-33. [PMID: 15187054]

389. Geersing GJ, Erkens PM, Lucassen WA, Bu¨ ller HR, Cate HT, Hoes AW, et al. Safe exclusion of pulmonary embolism using the Wells rule and qualitative d-dimer testing in primary care: prospective cohort study. BMJ. 2012;345:e6564. [PMID: 23036917]

390. Collins GS, Altman DG. Identifying patients with undetected gastro-oesophageal cancer in primary care: external validation of QCancer® (Gastro-Oesophageal). Eur J Cancer. 2013;49:1040-8. [PMID: 23159533]

391. de Vin T, Engels B, Gevaert T, Storme G, De Ridder M. Stereotactic radiotherapy for oligometastatic cancer: a prognostic model for survival. Ann Oncol. 2014;25:467-71. [PMID: 24355488]

392. Bernasconi P, Klersy C, Boni M, Cavigliano PM, Calatroni S, Giardini I, et al. World Health Organization classification in combination with cytogenetic markers improves the prognostic stratification of patients with de novo primary myelodysplastic syndromes. Br J Haematol. 2007;137:193-205. [PMID: 17408458]

393. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials. 1996;17:343-6. [PMID: 8889347]

394. Echouffo-Tcheugui JB, Woodward M, Kengne AP. Predicting a post-thrombolysis intracerebral hemorrhage: a systematic review. J Thromb Haemost. 2013;11:862-71. [PMID: 23469771]

395. Le Gal G, Righini M, Roy PM, Sanchez O, Aujesky D, Bounameaux H, et al. Prediction of pulmonary embolism in the emergency department: the revised Geneva score. Ann Intern Med. 2006;144:165-71. [PMID: 16461960]

396. Davis JL, Worodria W, Kisembo H, Metcalfe JZ, Cattamanchi A, Kawooya M, et al. Clinical and radiographic factors do not accurately diagnose smear-negative tuberculosis in HIV-infected inpatients in Uganda: a cross-sectional study. PLoS One. 2010;5:e9859. [PMID: 20361038]

397. Ji R, Shen H, Pan Y, Wang P, Liu G, Wang Y, et al; China National Stroke Registry (CNSR) Investigators. Risk score to predict gastrointestinal bleeding after acute ischemic stroke. BMC Gastroenterol. 2014;14:130. [PMID: 25059927]

398. Marrugat J, Subirana I, Ramos R, Vila J, Marin-Ibanez A, Guembe MJ, et al; FRESCO Investigators. Derivation and validation of a set of 10-year cardiovascular risk predictive functions in Spain: the FRESCO Study. Prev Med. 2014;61:66-74. [PMID: 24412897]

399. Hensgens MP, Dekkers OM, Goorhuis A, LeCessie S, Kuijper EJ. Predicting a complicated course of Clostridium difficile infection at the bedside. Clin Microbiol Infect. 2014;20:O301-8. [PMID: 24188103]

400. Hak E, Wei F, Nordin J, Mullooly J, Poblete S, Nichol KL. Development and validation of a clinical prediction rule for hospitalization due to pneumonia or influenza or death during influenza epidemics among community-dwelling elderly persons. J Infect Dis. 2004;189: 450-8. [PMID: 14745702]

401. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology. 2007;18:805-35. [PMID: 18049195]

402. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D’Agostino RB, et al. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. Lancet. 2009;373:739-45. [PMID: 19249635]

403. Lang TA, Altman DG. Basic statistical reporting for articles published in clinical medical journals: the SAMPL guidelines. In: Smart P, Maisonneuve H, Polderman A, eds. Science Editors' Handbook. European Association of Science Editors; 2013.

404. Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. Stat Med. 2013;32:2262-77. [PMID: 23034770]

405. Harrison DA, Parry GJ, Carpenter JR, Short A, Rowan K. A new risk prediction model for critical care: the Intensive Care National Audit & Research Centre (ICNARC) model. Crit Care Med. 2007;35:1091-8. [PMID: 17334248]

406. Brady AR, Harrison D, Black S, Jones S, Rowan K, Pearson G, et al. Assessment and optimization of mortality prediction tools for admissions to pediatric intensive care in the United Kingdom. Pediatrics. 2006;117:e733-42. [PMID: 16510615]

407. Kuijpers T, van der Windt DA, van der Heijden GJ, Twisk JW, Vergouwe Y, Bouter LM. A prediction rule for shoulder pain related sick leave: a prospective cohort study. BMC Musculoskelet Disord. 2006;7:97. [PMID: 17150087]

408. Pocock SJ, McCormack V, Gueyffier F, Boutitie F, Fagard RH, Boissel JP. A score for predicting risk of death from cardiovascular disease in adults with raised blood pressure, based on individual patient data from randomised controlled trials. BMJ. 2001;323:75-81. [PMID: 11451781]

409. Casikar I, Lu C, Reid S, Condous G. Prediction of successful expectant management of first trimester miscarriage: development and validation of a new mathematical model. Aust N Z J Obstet Gynaecol. 2013;53:58-63. [PMID: 23405997]

410. Godoy G, Chong KT, Cronin A, Vickers A, Laudone V, Touijer K, et al. Extent of pelvic lymph node dissection and the impact of standard template dissection on nomogram prediction of lymph node involvement. Eur Urol. 2011;60:195-201. [PMID: 21257258]

411. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis—an introduction to concepts and methods. Br J Cancer. 2003;89:431-6. [PMID: 12888808]

412. Wells P, Anderson D, Rodger M, Ginsberg J, Kearon C, Gent M, et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED d-dimer. Thromb Haemost. 2000;83:416-20. [PMID: 10744147]

413. Cole TJ. Scaling and rounding regression-coefficients to integers. Appl Stat. 1993;42:261-8.

414. Sullivan LM, Massaro JM, D’Agostino RB Sr. Presentation of multivariate data for clinical use: the Framingham study risk score functions. Stat Med. 2004;23:1631-60. [PMID: 15122742]

415. Moons KG, Harrell FE, Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? J Clin Epidemiol. 2002;55:1054-5. [PMID: 12464384]

416. Nijman RG, Vergouwe Y, Thompson M, van Veen M, van Meurs AH, van der Lei J, et al. Clinical prediction model to aid emergency doctors managing febrile children at risk of serious bacterial infections: diagnostic study. BMJ. 2013;346:f1706. [PMID: 23550046]

417. Royston P, Altman DG. Visualizing and assessing discrimination in the logistic regression model. Stat Med. 2010;29:2508-20. [PMID: 20641144]

418. Tas¸ U, Steyerberg EW, Bierma-Zeinstra SM, Hofman A, Koes BW, Verhagen AP. Age, gender and disability predict future disability in older people: the Rotterdam Study. BMC Geriatrics. 2011;11:22. [PMID: 21569279]

419. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148:839-43. [PMID: 6878708]

420. Pencina MJ, D’Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11-21. [PMID: 21204120]

421. Pepe MS, Janes H. Reporting standards are needed for evaluations of risk reclassification. Int J Epidemiol. 2011;40:1106-8. [PMID:21571811]

422. Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. Semin Oncol. 2010;37:31-8. [PMID: 20172362]

423. Sanders MS, de Jonge RC, Terwee CB, Heymans MW, Koomen I, Ouburg S, et al. Addition of host genetic variants in a prediction rule for post meningitis hearing loss in childhood: a model updating study. BMC Infect Dis. 2013;13:340. [PMID: 23879305]

424. Kramer AA, Zimmerman JE. A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. BMC Med Inform Decis Mak. 2010;10:27. [PMID: 20465830]

425. Neely D, Feinglass J, Wallace WH. Developing a predictive model to assess applicants to an internal medicine residency. J Grad Med Educ. 2010;2:129-32. [PMID: 21975899]

426. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. J Clin Epidemiol. 2007;60:324-9. [PMID: 17346604]

427. Horton R. The hidden research paper. JAMA. 2002;287:2775-8. [PMID: 12038909]

428. Docherty M, Smith R. The case for structuring the discussion of cientific papers. BMJ. 1999;318:1224-5. [PMID: 10231230]

429. Ioannidis JP. Research needs grants, funding and money—missing something? Eur J Clin Invest. 2012;42:349-51. [PMID: 22050119]

430. Janssens AC, Ioannidis JP, Bedrosian S, Boffetta P, Dolan SM, Dowling N, et al. Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. Eur J Clin Invest. 2011;41:1010-35. [PMID: 21434890]

431. Collins GS. Cardiovascular disease risk prediction in the UK. Primary Care Cardiovascular Journal. 2013;6:125-8.

432. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. BMJ. 2009;339:b2584. [PMID: 19584409]

433. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. BMJ. 2010;340:c2442. [PMID: 20466793]

434. Perry JJ, Sharma M, Sivilotti ML, Sutherland J, Symington C, Worster A, et al. Prospective validation of the ABCD2 score for patients in the emergency department with transient ischemic attack. CMAJ. 2011;183:1137-45. [PMID: 21646462]

435. Clarke M, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals: islands in search of continents? JAMA. 1998;280:280-2. [PMID: 9676682]

436. Ioannidis JP, Polyzos NP, Trikalinos TA. Selective discussion and transparency in microarray research findings for cancer outcomes. Eur J Cancer. 2007;43:1999-2010. [PMID: 17629475]

437. Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? Anesth Analg. 2005;100:1525-32. [PMID: 15845719]

438. Kappen TH, Moons KG, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, van Klei WA. Impact of risk assessments on prophylactic antiemetic prescription and the incidence of postoperative nausea and vomiting: a cluster-randomized trial. Anesthesiology. 2014;120:343-54. [PMID: 24105403]

439. Poldervaart JM, Reitsma JB, Koffijberg H, Backus BE, Six AJ, Doevendands PA, et al. The impact of the HEART risk score in the early assessment of patients with acute chest pain: design of a stepped wedge, cluster randomised trial. BMC Cardiovasc Disord. 2013;13:77. [PMID: 24070098]

440. Hutchings HA, Evans BA, Fitzsimmons D, Harrison J, Heaven M, Huxley P, et al. Predictive risk stratification model: a progressive cluster-randomised trial in chronic conditions management (PRISMATIC) research protocol. Trials. 2013;14:301. [PMID: 24330749]

441. Ioannidis JP. More than a billion people taking statins? Potential implications of the new cardiovascular guidelines. JAMA. 2014;311:463-4. [PMID: 24296612]

442. Ioannidis JP, Tzoulaki I. What makes a good predictor? The evidence applied to coronary artery calcium score. JAMA. 2010;303:1646-7. [PMID: 20424257]

443. Mrdovic I, Savic L, Krljanac G, Asanin M, Perunicic J, Lasica R, et al. Predicting 30-day major adverse cardiovascular events after primary percutaneous coronary intervention. The RISK-PCI score. Int J Cardiol. 2013;162:220-7. [PMID: 21663982]

444. Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. Circulation. 2008;118:2243-51. [PMID: 18997194]

445. World Medical Association. Declaration of Geneva. Accessed at www.wma.net/en/30publications/10policies/g1/ on 24 June 2008.

446. Council for International Organizations of Medical Sciences. International ethical guidelines for biomedical research involving human subjects. Bull Med Ethics. 2002;182:17-23. [PMID: 14983848]

447. Arnold DH, Gebretsadik T, Abramo TJ, Sheller JR, Resha DJ, Hartert TV. The Acute Asthma Severity Assessment Protocol (AASAP) study: objectives and methods of a study to develop an acute asthma clinical prediction rule. Emerg Med J. 2012;29:444-50. [PMID: 21586757]

448. Azagra R, Roca G, Encabo G, Prieto D, Aguye A, Zwart M, et al. Prediction of absolute risk of fragility fracture at 10 years in a Spanish population: validation of the WHO FRAX tool in Spain. BMC Musculoskelet Disord. 2011;12:30. [PMID: 21272372]

449. Collins SP, Lindsell CJ, Jenkins CA, Harrell FE, Fermann GJ, Miller KF, et al. Risk stratification in acute heart failure: rationale and design of the STRATIFY and DECIDE studies. Am Heart J. 2012;164:825-34. [PMID: 23194482]

450. Hafkamp-de Groen E, Lingsma HF, Caudri D, Wijga A, Jaddoe VW, Steyerberg EW, et al. Predicting asthma in preschool children with asthma symptoms: study rationale and design. BMC Pulm Med. 2012;12:65. [PMID: 23067313]

451. Hess EP, Wells GA, Jaffe A, Stiell IG. A study to derive a clinical decision rule for triage of emergency department patients with chest pain: design and methodology. BMC Emerg Med. 2008;8:3. [PMID:18254973]

452. Horisberger T, Harbarth S, Nadal D, Baenziger O, Fischer JE. G-CSF and IL-8 for early diagnosis of sepsis in neonates and critically ill children—safety and cost effectiveness of a new laboratory prediction model: study protocol of a randomized controlled trial [ISRCTN91123847]. Crit Care. 2004;8:R443-50. [PMID: 15566590]

453. Liman TG, Zietemann V, Wiedmann S, Jungehuelsing GJ, Endres M, Wollenweber FA, et al. Prediction of vascular risk after stroke—protocol and pilot data of the Prospective Cohort with Incident Stroke (PROSCIS). Int J Stroke. 2013;8:484-90. [PMID:22928669]

454. Mann DM, Kannry JL, Edonyabo D, Li AC, Arciniega J, Stulman J, et al. Rationale, design, and implementation protocol of an electronic health record integrated clinical prediction rule (iCPR) randomized trial in primary care. Implement Sci. 2011;6:109. [PMID:21929769]

455. Meijs MF, Bots ML, Vonken EJ, Cramer MJ, Melman PG, Velthuis BK, et al. Rationale and design of the SMART Heart study: a prediction model for left ventricular hypertrophy in hypertension. Neth Heart J. 2007;15:295-8. [PMID: 18030317]

456. Mrdovic I, Savic L, Perunicic J, Asanin M, Lasica R, Marinkovic J, et al. Development and validation of a risk scoring model to predict net adverse cardiovascular outcomes after primary percutaneous coronary intervention in patients pretreated with 600 mg clopidogrel: rationale and design of the RISK-PCI study. J Interv Cardiol. 2009;22:320-8. [PMID: 19515084]

457. Nee RJ, Vicenzino B, Jull GA, Cleland JA, Coppieters MW. A novel protocol to develop a prediction model that identifies patients with nerve-related neck and arm pain who benefit from the early introduction of neural tissue management. Contemp Clin Trials. 2011;32:760-70. [PMID: 21718803]

458. Pita-Ferna´ndez S, Pe´ rtega-Diaz S, Valde´ s-Can˜edo F, Seijo-Bestilleiro R, Seoane-Pillado T, Ferna´ ndez-Rivera C, et al. Incidence of cardiovascular events after kidney transplantation and cardiovascular risk scores: study protocol. BMC Cardiovasc Disord. 2011;11:2. [PMID: 21639867]

459. Sanfelix-Genoves J, Peiro S, Sanfelix-Gimeno G, Giner V, Gil V, Pascual M, et al. Development and validation of a population-based prediction scale for osteoporotic fracture in the region of Valencia, Spain: the ESOSVAL-R study. BMC Public Health. 2010;10:153. [PMID: 20334639]

460. Siebeling L, ter Riet G, van der Wal WM, Geskus RB, Zoller M, Muggensturm P, et al. ICE COLD ERIC—International collaborative effort on chronic obstructive lung disease: exacerbation risk index cohorts—study protocol for an international COPD cohort study. BMC Pulm Med. 2009;9:15. [PMID: 19419546]

461. Canadian CT Head and C-Spine (CCC) Study Group. Canadian C-Spine Rule study for alert and stable trauma patients: I. Background and rationale. CJEM. 2002;4:84-90. [PMID: 17612425]

462. Canadian CT Head and C-Spine (CCC) Study Group. Canadian C-Spine Rule study for alert and stable trauma patients: II. Study objectives and methodology. CMAJ. 2002;4:185-93. [PMID: 17609004]

463. van Wonderen KE, van der Mark LB, Mohrs J, Geskus RB, van der Wal WM, van Aalderen WM, et al. Prediction and treatment of asthma in preschool children at risk: study design and baseline data of a prospective cohort study in general practice (ARCADE). BMC Pulm Med. 2009;9:13. [PMID: 19368704]

464. Waldron CA, Gallacher J, van der Weijden T, Newcombe R, Elwyn G. The effect of different cardiovascular risk presentation formats on intentions, understanding and emotional affect: a randomized controlled trial using a web-based risk formatter (protocol). BMC Med Inform Decis Mak. 2010;10:41. [PMID: 20673347]

465. Laine C, Guallar E, Mulrow C, Taichman DB, Cornell JE, Cotton D, et al. Closing in on the truth about recombinant human bone morphogenetic protein-2: evidence synthesis, data sharing, peer review, and reproducible research. Ann Intern Med. 2013;158:916-8. [PMID: 23778911]

466. Peng RD. Reproducible research and Biostatistics. Biostatistics. 2009;10:405-8. [PMID: 19535325]

467. Keiding N. Reproducible research and the substantive context. Biostatistics. 2010;11:376-8. [PMID: 20498225]

468. Vickers AJ. Whose data set is it anyway? Sharing raw data from randomized trials. Trials. 2006;7:15. [PMID: 16704733]

469. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Jones DR, Heney D, et al. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. Br J Cancer. 2003;88:1191-8. [PMID: 12698183]

470. Riley RD, Sauerbrei W, Altman DG. Prognostic markers in cancer: the evolution of evidence from single studies to meta-analysis, and beyond. Br J Cancer. 2009;100:1219-29. [PMID: 19367280]

471. Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol. 2007;60:431-9. [PMID: 17419953]

472. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. BMJ. 2009;339:b4184. [PMID: 20042483]

473. Groves T. BMJ policy on data sharing. BMJ. 2010;340:c564. [PMID: 20110310]

474. Marchionni L, Afsari B, Geman D, Leek JT. A simple and reproducible breast cancer prognostic test. BMC Genomics. 2013;14:336. [PMID: 23682826]

475. Loder E, Groves T, Macauley D. Registration of observational studies. BMJ. 2010;340:c950. [PMID: 20167643]

476. Chavers S, Fife D, Wacholtz M, Stang P, Berlin J. Registration of Observational Studies: perspectives from an industry-based epidemiology group. Pharmacoepidemiol Drug Saf. 2011;20:1009-13. [PMID: 21953845]

477. Should protocols for observational studies be registered? Lancet. 2010;375:348. [PMID: 20113809]

478. Altman DG. The time has come to register diagnostic and prognostic research. Clin Chem. 2014;60:580-2. [PMID: 24520099]

479. The registration of observational studies—when metaphors go bad. Epidemiology. 2010;21:607-9. [PMID: 20657291]

480. Sørensen HT, Rothman KJ. The prognosis of research. BMJ. 2010;340:c703. [PMID: 20164129]

481. Vandenbroucke JP. Registering observational research: second thoughts. Lancet. 2010;375:982-3. [PMID: 20304239]

482. Williams RJ, Tse T, Harlan WR, Zarin DA. Registration of observational studies: Is it time? CMAJ. 2010;182:1638-42. [PMID: 20643833]

483. Lenzer J. Majority of panelists on controversial new cholesterol guideline have current or recent ties to drug manufacturers. BMJ. 2013;347:f6989. [PMID: 24264770]

484. Lenzer J, Hoffman JR, Furberg CD, Ioannidis JP ; Guideline Panel Review Working Group. Ensuring the integrity of clinical practice guidelines: a tool for protecting patients. BMJ. 2013;347:f5535. [PMID: 24046286]

485. Simera I. Get the content right: following reporting guidelines will make your research paper more complete, transparent and usable. J Pak Med Assoc. 2013;63:283-5. [PMID: 23894916]

486. Simera I, Kirtley S, Altman DG. Reporting clinical research: guidance to encourage accurate and transparent research reporting. Maturitas. 2012;72:84-7. [PMID: 22440533]

487. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. BMC Med. 2010;8:24. [PMID: 20420659]

488. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151:264-9. [PMID: 19622511]

489. Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, et al; STrengthening the REporting of Genetic Association Studies. STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. PLoS Med. 2009;6:e22. [PMID: 19192942]

490. Kilkenny C, Browne W, Cuthill IC, Emerson M, Altman DG; NC3Rs Reporting Guidelines Working Group. Animal research: reporting in vivo experiments: the ARRIVE guidelines. J Gene Med. 2010;12:561-3. [PMID: 20607692]

491. Gagnier JJ, Kienle G, Altman DG, Moher D, Sox H, Riley D; CARE Group. The CARE guidelines: consensus-based clinical case reporting guideline development. J Med Case Rep. 2013;7:223. [PMID: 24228906]

492. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol. 2010;10:7. [PMID: 20085642]

493. Little RJ, Rubin DB. Statistical Analysis With Missing Data. Hoboken, NJ: Wiley; 2002.

494. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons; 1987.

495. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30:377-99. [PMID: 21225900]

496. Harel O, Pellowski J, Kalichman S. Are we missing the importance of missing values in HIV prevention randomized clinical trials? Review and recommendations. AIDS Behav. 2012;16:1382-93. [PMID: 22223301]

497. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8:3-15. [PMID: 10347857]

498. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9:57. [PMID: 19638200]

499. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18:681-94. [PMID: 10204197]

500. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Stat Med. 2008;27:3227-46. [PMID: 18203127]

501. Turner EL, Dobson JE, Pocock SJ. Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. Epidemiol Perspect Innov. 2010;7:9.

502. van Walraven C, Hart RG. Leave ‘em alone—why continuous variables should be analyzed as such. Neuroepidemiology. 2008;30:138-9. [PMID: 18421216]

503. Vickers AJ, Lilja H. Cutpoints in clinical chemistry: time for fundamental reassessment. Clin Chem. 2009;55:15-7. [PMID: 19028819]

504. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol. 2012;12:21. [PMID: 22375553]

505. Dawson NV, Weiss R. Dichotomizing continuous variables in statistical analysis: a practice to avoid. Med Decis Making. 2012;32:225-6. [PMID: 22457338]

506. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat. 1994;43:429-67.

507. Harrell FE Jr, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst. 1988;80:1198-202. [PMID: 3047407]

508. Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. Bioinformatics. 2007;23:1768-74. [PMID: 17485430]

509. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst. 2010;102:464-74. [PMID: 20233996]

510. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 2007;99:147-57. [PMID:17227998]

511. Boulesteix AL. Validation in bioinformatics and molecular medicine. Brief Bioinform. 2011;12:187-8. [PMID: 21546447]

512. Jelizarow M, Guillemot V, Tenenhaus A, Strimmer K, Boulesteix AL. Over-optimism in bioinformatics: an illustration. Bioinformatics. 2010;26:1990-8. [PMID: 20581402]

513. Vickers AJ, Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). Urology. 2010;76:1298-301. [PMID: 21030068]

514. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33:517-35. [PMID: 24002997]

515. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Methods Med Res. 2014 Apr 7. [Epub ahead of print]. [PMID: 23907781]

516. Vach W. Calibration of clinical prediction rules does not just assess bias. J Clin Epidemiol. 2013;66:1296-301. [PMID: 24021610]

517. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic-regression models. Stat Med. 1991;10:1213-26. [PMID: 1925153]

518. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45:562-5.

519. D’Agostino RB, Nam BH. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Balakrishnan N, Rao CR, eds. Handbook of Statistics, Survival Methods. Amsterdam: Elsevier; 2004:1-25.

520. Grønnesby JK, Borgan O. A method for checking regression models in survival analysis based on the risk score. Lifetime Data Anal. 1996;2:315-28. [PMID: 9384628]

521. Bertolini G, D’Amico R, Nardi D, Tinazzi A, Apolone G. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. J Epidemiol Biostat. 2000;5:251-3. [PMID: 11055275]

522. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. Crit Care Med. 2007;35:2052-6. [PMID: 17568333]

523. Marcin JP, Romano PS. Size matters to a model's fit. Crit Care Med. 2007;35:2212-3. [PMID: 17713369]

524. Bannister CA, Poole CD, Jenkins-Jones S, Morgan CL, Elwyn G, Spasic I, et al. External validation of the UKPDS risk engine in incident type 2 diabetes: a need for new type 2 diabetes-specific risk equations. Diab Care. 2014;37:537-45. [PMID: 24089541]

525. Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW, Van Calster B. Assessing calibration of multinomial risk prediction models. Stat Med. 2014;33:2585-96. [PMID: 24549725]

526. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. 2008;54:17-23. [PMID: 18024533]

527. Pencina MJ, D’Agostino RB Sr, Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. Stat Med. 2012;31:1543-53. [PMID: 22344892]

528. Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the c-statistic to nominal polytomous outcomes: the polytomous discrimination index. Stat Med. 2012;31:2610-26. [PMID: 22733650]

529. Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics. 2014;15:526-39. [PMID: 24493091]

530. Pencina MJ, D’Agostino RB Sr, Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. Stat Med. 2012;31:101-13. [PMID: 22147389]

531. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part III: multivariate data analysis—choosing a model and assessing its adequacy and fit. Br J Cancer. 2003;89:605-11. [PMID: 12915864]

532. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for the systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11:e1001744. [PMID: 25314315]