DOI: https://doi.org/10.17816/DD110794

Прозрачная отчётность о многофакторной предсказательной модели для индивидуального прогнозирования или диагностики (TRIPOD): разъяснения и уточнения

K.G.M. Moons[1], D.G. Altman[2], J.B. Reitsma[1], J.P.A. Loannidis[3], P. Macaskill[4], E.W. Steyerberg[5], A.J. Vickers[6], D.F. Ransohoff[7], G.S. Collins[2]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, Нидерланды
[2] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford OX3 7LD, Великобритания
[3] Stanford Prevention Research Center, School of Medicine, Stanford University, 291 Campus Drive, Room LK3C02, Li Ka Shing Building, 3rd Floor, Stanford, CA 943055101, США
[4] Screening & Test Evaluation Program (STEP), School of Public Health, Edward Ford Building (A27), Sydney Medical School, University of Sydney, Sydney, NSW 2006, Австралия
[5] Department of Public Health, Erasmus MC-University Medical Center Rotterdam, PO Box 2040, 3000 CA, Rotterdam, Нидерланды
[6] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 307 East 63rd Street, 2nd Floor, Box 44, New York, NY 10065, США
[7] Departments of Medicine and Epidemiology, University of North Carolina at Chapel Hill, 4103 Bioinformatics, CB 7080, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7080, США

АННОТАЦИЯ

Руководство TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) содержит контрольный перечень из 22 пунктов рекомендаций, предложенных для повышения качества отчётности по исследованиям, в которых разрабатывали, проверяли или обновляли предсказательные модели для диагностики или прогнозирования. Руководство TRIPOD направлено на повышение прозрачности отчёта об исследовании предсказательной модели, независимо от использованных методов. Этот документ с пояснениями и уточнениями включает обоснование, разъяснения значений каждого пункта рекомендаций, обсуждение важности прозрачной отчётности для оценки риска систематических ошибок и клинической полезности предсказательной модели. Каждая рекомендация руководства TRIPOD подробно объясняется, приводятся опубликованные примеры правильного представления результатов. Документ также содержит ценную справочную информацию, которую следует учитывать при разработке, проведении и анализе исследований предсказательных моделей. Рекомендуем авторам включать в свои работы все пункты контрольного перечня, что облегчит оценку исследования редакторами, рецензентами, читателями и исследователями, проводящими систематическое обобщение результатов таких исследований. Контрольный перечень TRIPOD также доступен по адресу: www.tripod-statement.org.

Информацию о членах инициативной группы TRIPOD см. в **Приложении**.

Данная статья является переводом на русский язык. Оригинальная статья опубликована в *Annals of Internal Medicine.* 2015;162(1):W1–W73. doi: 10.7326/M14-0698. Перевод и повторная публикация осуществлены с разрешения правообладателя. Перевод и научное редактирование выполнены д.м.н. Р.Т. Сайгитовым (ORCID: 0000-0002-8915-6153).

Как цитировать

Moons K.G.M., Altman D.G., Reitsma J.B., Loannidis J.P.A., Macaskill P., Steyerberg E.W., Vickers A.J., Ransohoff D.F., Collins G.S. Прозрачная отчётность о многофакторной предсказательной модели для индивидуального прогнозирования или диагностики (TRIPOD): разъяснения и уточнения // *Digital Diagnostic*. 2022. Т. 3, №3. С. 232–322. DOI: https://doi.org/10.17816/DD110794

Рукопись получена: 15.12.2021 **Рукопись одобрена:** 10.08.2022 **Опубликована:** 05.10.2022

DOI: https://doi.org/10.17816/DD110794

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Translation in to Russian

Karel G.M. Moons[1], Douglas G. Altman[2], Johannes B. Reitsma[1], John P.A. Loannidis[3], Petra Macaskill[4], Ewout W. Steyerberg[5], Andrew J. Vickers[6], David F. Ransohoff[7], Gary S. Collins[2]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands
[2] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford OX3 7LD, United Kingdom
[3] Stanford Prevention Research Center, School of Medicine, Stanford University, 291 Campus Drive, Room LK3C02, Li Ka Shing Building, 3rd Floor, Stanford, CA 943055101, USA
[4] Screening & Test Evaluation Program (STEP), School of Public Health, Edward Ford Building (A27), Sydney Medical School, University of Sydney, Sydney, NSW 2006, Australia
[5] Department of Public Health, Erasmus MC-University Medical Center Rotterdam, PO Box 2040, 3000 CA, Rotterdam, the Netherlands
[6] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 307 East 63rd Street, 2nd Floor, Box 44, New York, NY 10065, USA
[7] Departments of Medicine and Epidemiology, University of North Carolina at Chapel Hill, 4103 Bioinformatics, CB 7080, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7080, USA

ABSTRACT

The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement includes a 22-item checklist, which aims to improve the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes. The TRIPOD Statement aims to improve the transparency of the reporting of a prediction model study regardless of the study methods used. This explanation and elaboration document describes the rationale; clarifies the meaning of each item; and discusses why transparent reporting is important, with a view to assessing risk of bias and clinical usefulness of the prediction model. Each checklist item of the TRIPOD Statement is explained in detail and accompanied by published examples of good reporting. The document also provides a valuable reference of issues to consider when designing, conducting, and analyzing prediction model studies. To aid the editorial process and help peer reviewers and, ultimately, readers and systematic reviewers of prediction model studies, it is recommended that authors include a completed checklist in their submission. The TRIPOD checklist can also be downloaded from *www.tripod-statement.org*.

For members of the TRIPOD Group, see the Appendix.

This article is the translation in to Russian by Dr. Ruslan Saygitov (ORCID: 0000-0002-8915-6153) from the original published in [*Ann Intern Med*. 2015; 162:W1-W73. doi: 10.7326/M14-0698].

To cite this article

Moons KGM, Altman DG, Reitsma JB, Loannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Translation in to Russian. *Digital Diagnostic*. 2022;3(3):232–322. DOI: https://doi.org/10.17816/DD110794

Received: 15.12.2021 **Accepted:** 10.08.2022 **Published:** 05.10.2022

DOI: https://doi.org/10.17816/DD110794

个体预后或诊断的多变量预测模型透明报告(TRIPOD)：解释和说明

Karel G.M. Moons[1], Douglas G. Altman[2], Johannes B. Reitsma[1], John P.A. Loannidis[3], Petra Macaskill[4], Ewout W. Steyerberg[5], Andrew J. Vickers[6], David F. Ransohoff[7], Gary S. Collins[2]

[1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands
[2] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Oxford OX3 7LD, United Kingdom
[3] Stanford Prevention Research Center, School of Medicine, Stanford University, 291 Campus Drive, Room LK3C02, Li Ka Shing Building, 3rd Floor, Stanford, CA 943055101, USA
[4] Screening & Test Evaluation Program (STEP), School of Public Health, Edward Ford Building (A27), Sydney Medical School, University of Sydney, Sydney, NSW 2006, Australia
[5] Department of Public Health, Erasmus MC-University Medical Center Rotterdam, PO Box 2040, 3000 CA, Rotterdam, the Netherlands
[6] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 307 East 63rd Street, 2nd Floor, Box 44, New York, NY 10065, USA
[7] Departments of Medicine and Epidemiology, University of North Carolina at Chapel Hill, 4103 Bioinformatics, CB 7080, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7080, USA

简评

TRIPOD（Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis）指南包含22项建议的清单（checklist），以提高诊断或预测模型(prediction model)已被开发、测试或更新的研究报告的质量。TRIPOD指南旨在提高预测模型研究报告的透明度，无论使用何种方法。该文件含有解释和说明，包括理由、对建议每一点含义的澄清、对透明报告对评估偏倚风险的重要性的讨论以及预测模型的临床有用性。TRIPOD指南中的每项建议都有详细的解释，并附有已发布的正确报告结果的示例。本文还提供了在设计、进行和分析预测模型研究时应考虑的有价值的参考信息。我们建议作者在他们的论文中包含所有清单项目，这将有助于编辑、审稿人、读者和对此类研究进行系统综合的研究人员对研究进行评估。TRIPOD清单也可在以下网站获得：www.tripod-statement.org

本文为俄文译本。原文发表在Annals of Internal Medicine. 2015;162(1):W1－W73. doi: 10.7326/M14-0698.经版权所有者许可，翻译并重新出版。

To cite this article
Moons KGM, Altman DG, Reitsma JB, Loannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. 个体预后或诊断的多变量预测模型透明报告(TRIPOD)：解释和说明. Translation in to Russian. *Digital Diagnostic.* 2022;3(3):232–322. DOI: https://doi.org/10.17816/DD110794

收到： 15.12.2021 **接受：** 10.08.2022 **发布日期：** 05.10.2022

ВВЕДЕНИЕ

В медицине поставщики медицинских услуг принимают многочисленные решения (зачастую совместно с пациентом) на основе оцениваемой вероятности наличия определённого заболевания или состояния (диагностические условия, *diagnostic setting*) или определённого события, которое произойдёт в будущем (прогностические условия, *prognostic setting*) у человека. В диагностике вероятность наличия определённого заболевания может использоваться, например, для информирования о необходимости дальнейшего обследования пациентов, начала лечения или убеждения пациентов в том, что серьёзная причина их симптомов маловероятна. В прогнозировании предсказания могут использоваться для планирования образа жизни или терапевтических решений на основе риска наступления определённого исхода или состояния здоровья в течение определённого периода [1–3]. Такие оценки риска могут быть полезны при распределении пациентов по группам риска в исследованиях, посвящённых терапевтическим вмешательствам [4–7].

И в диагностике, и в прогнозировании оценки вероятности обычно основываются на комбинировании информации о многочисленных предикторах (*predictors*), наблюдаемых или измеренных у человека [1, 2, 8–10]. Единичные предикторы, как правило, не дают надёжных оценок диагностической или прогностической вероятности или рисков [8, 11]. Практически во всех областях медицины многофакторные диагностические и прогностические модели, предсказывающие риск, разрабатывают, проверяют, обновляют и внедряют, чтобы помочь врачам и отдельным лицам в оценке вероятностей развития заболевания и повлиять на принятие ими решений.

Многофакторная предсказательная модель (*multivariable prediction model*) — математическое уравнение, связывающее несколько предикторов у отдельного индивида с вероятностью или риском наличия (диагнозом) или возможным возникновением в будущем (прогнозом) конкретного исхода [10, 12]. Предсказательную модель (*prediction model*) ещё называют моделью предсказания рисков (*risk prediction model*), предиктивной моделью (*predictive model*), прогностическим индексом (*prognostic index*) или правилом (*prognostic rule*), шкалой риска (*risk score*) [9].

Предикторы называют также ковариатами (*covariates*), маркерами риска (*risk indicators*), прогностическими факторами (*prognostic factors*), детерминантами (*determinants*), результатами тестирования (*test results*) или, в статистическом смысле, независимыми переменными (*independent variables*). Предикторами могут быть демографические характеристики (например, возраст и пол), данные анамнеза, физикального обследования, инструментальной визуализации, электрофизиологического исследования, анализов крови и мочи, цитологического и гистологического исследований, стадии или характеристики болезни, результаты геномных, протеомных, транскриптомных, фармакогеномных, метаболомных и других новых биологических измерений.

ДИАГНОСТИЧЕСКИЕ И ПРОГНОСТИЧЕСКИЕ ПРЕДСКАЗАТЕЛЬНЫЕ МОДЕЛИ

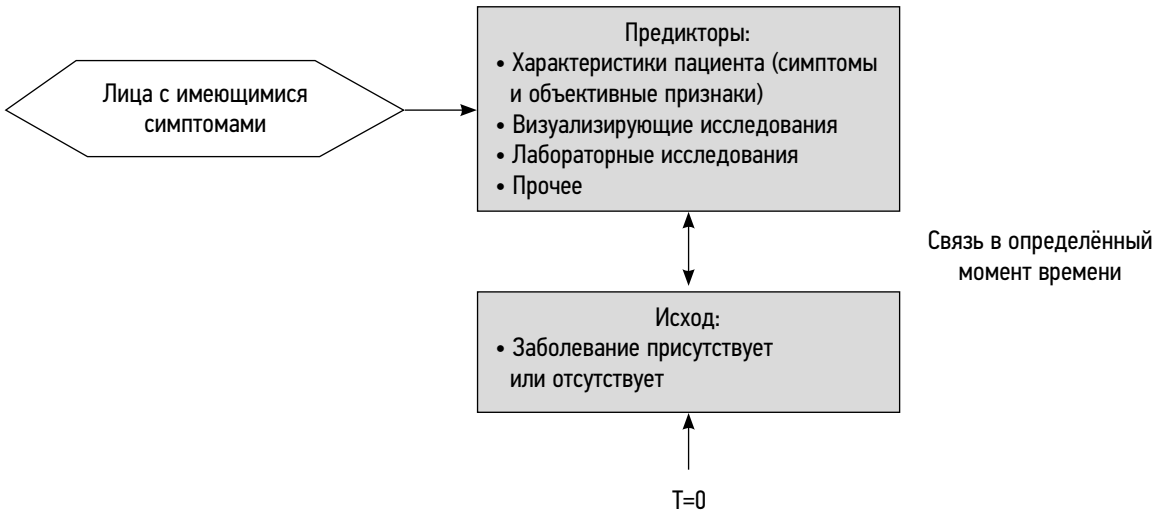
Многофакторные предсказательные модели делятся на две широкие категории: диагностические и прогностические (**вставка А**). В диагностической модели (*diagnostic model*) несколько (2 и более) предикторов, часто называемых результатами диагностических тестов (*diagnostic test results*), объединяют для оценки вероятности того, что определённое состояние или заболевание присутствует (или отсутствует) в момент предсказания (**вставка Б**). Такие модели разрабатывают и впоследствии применяют в отношении лиц, у которых подозревают такое состояние.

В прогностической модели (*prognostic model*) несколько предикторов объединяют для оценки вероятности конкретного исхода или события (например, наступления смерти, развития рецидива болезни, осложнения или ответа на терапию), которое может произойти в конкретный период в будущем. Этот период может варьировать от нескольких часов (например, предсказание послеоперационных осложнений [13]) до нескольких недель или месяцев (например, предсказание 30-суточной летальности после операции на сердце [14]) или даже лет (например, 5-летний риск развития сахарного диабета 2-го типа [15]).

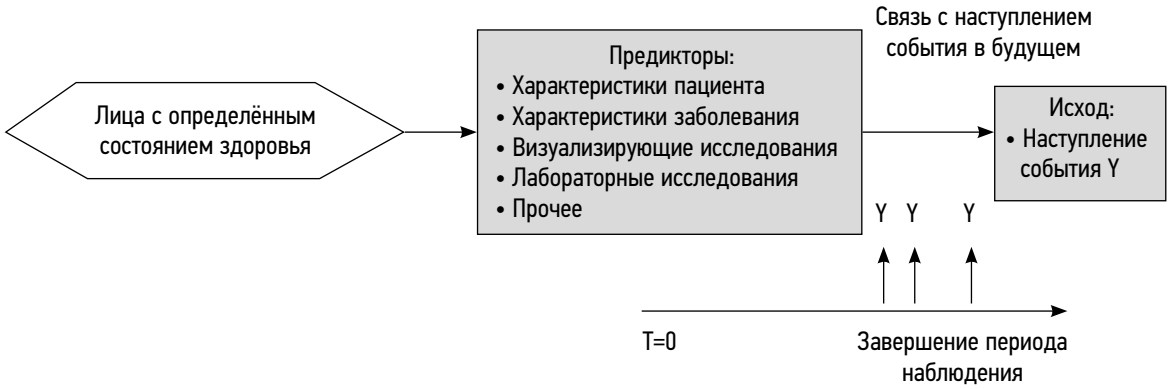
Суть предсказания в диагностике заключается в оценке вероятности того, что конкретный исход или заболевание присутствует (или нет) у индивидуума в определённый момент времени, т.е. в момент предсказания (T=0). В прогнозировании предсказание даётся в отношение того, наступит ли конкретное событие или исход у индивидуума в течение определённого периода времени. Иными словами, в диагностическом предсказании предполагается одномоментная связь (*cross-sectional relationship*), тогда как в прогностическом — отсроченная связь (*longitudinal relationship*). Тем не менее в исследованиях диагностических моделей часто необходим временно́й интервал между измерением предиктора (проведением исследуемого теста, *index test*) и выполнением референсного теста (*reference standard*). В идеале этот интервал должен быть как можно короче, без начала какого-либо лечения в этот период.

Прогностические модели разрабатываются и применяются в отношении лиц, подверженных риску наступления определённого исхода, пациентов с уже выявленным заболеванием или здоровых лиц. Например, это могут быть модели для предсказания

Исследование многофакторной диагностической модели



Исследование многофакторной прогностической модели



Вставка А. Схематическое представление исследований диагностических и прогностических предсказательных моделей

Несмотря на различия в природе предсказания (отношение ко времени), между диагностическими и прогностическими моделями есть много общего, в том числе:

- Тип исхода часто двоичен (*binary*): целевое заболевание присутствует или отсутствует (при диагностике); событие в будущем возникает или не возникает (при прогнозе).
- Основная цель — с учётом значений двух и более предикторов оценить вероятность (*probability*) наличия или наступления целевого состояния у пациентов для их последующего информирования и принятия клинического решения.
- Одинаковые проблемы, свойственные разработке многофакторной предсказательной модели, а именно выбор предикторов, стратегии построения моделей, обработка предикторов с непрерывной шкалой измерения (*continuous predictors*), опасность чрезмерной аппроксимации или переобучения (*overfitting*).
- Одинаковые показатели для оценки эффективности модели (*model performance*).

Ниже приведены различные термины для обозначения сходных характеристик исследований диагностического и прогностического моделирования.

Исследование моделирования диагностического предсказания (*diagnostic prediction modeling study*)	Частичная проверка (*partial verification*)	Прогностические факторы или показатели (*prognostic factors / indicators*)
Исследуемые диагностические тесты (*index tests*)	*Объясняющие переменные (explanatory variables), предикторы, ковариаты (X-переменные)*	Событие (event; наступление в будущем: да или нет)
Целевое (*target*) заболевание / патология (наличие или отсутствие)	*Исход, результат (Y-переменная, outcome)*	Определение и регистрация события (*event measurement*)
Референсный диагностический тест (*reference standard*) и подтверждение диагноза (*disease verification*)	*Отсутствующие исходы (missing outcomes)*	Выбывшие из-под наблюдения (*loss to follow-up*) и цензурирование (*censoring*)
	Исследование моделирования прогностического предсказания (*prognostic prediction modeling study*)	

Вставка Б. Сходства и различия между диагностическими и прогностическими предсказательными моделями

возникновения рецидивов, осложнений или наступления смерти в определённый период после установления конкретного диагноза. Но также это могут быть модели для предсказания наступления исхода в течение определённого периода у лиц без конкретного заболевания, например, в случае предсказания риска развития диабета 2-го типа [16], сердечно-сосудистых событий у здоровых людей среднего возраста [17], риска развития преэклампсии у беременных [18]. Таким образом, мы используем термин «прогностический» (*prognostic*) в широком смысле, имея в виду предсказание исхода в будущем у лиц, подверженных риску этого исхода, а не в узком смысле как предсказание исхода у пациентов с определённым заболеванием, получающих или не получающих лечение [1].

Основное различие между диагностическими и прогностическими моделями заключается в концепции времени. Исследования диагностических моделей, как правило, одномоментные (*cross-sectional*), а прогностических моделей — когортные (*longitudinal*). В настоящей статье и диагностические, и прогностические модели мы называем предсказательными моделями (*prediction models*), уделяя внимание вопросам, свойственным каждому типу моделей.

РАЗРАБОТКА, ПРОВЕРКА И ОБНОВЛЕНИЕ ПРЕДСКАЗАТЕЛЬНЫХ МОДЕЛЕЙ

Исследования предсказательных моделей могут быть посвящены разработке (*development*) новой модели [10], оценке характеристик модели (часто называемой проверкой модели, *model validation*) с последующим её обновлением (*updating*) или без него [19–21] или в комбинации выше перечисленного (**вставка В** и **рис. 1**).

Разработка предсказательной модели включает выбор предикторов и их объединение в многофакторную модель. Для предсказания одномоментных (диагностических) и краткосрочных исходов (например, 30-суточной летальности) обычно используют логистическую регрессию (*logistic regression*), для предсказания долгосрочных исходов (например, 10-летнего риска) — регрессию Кокса (*Cox regression*). Исследования предсказательных моделей могут быть нацелены также и на количественную оценку дополнительной предсказательной ценности (*incremental / added predictive value)* конкретного предиктора (например, недавно обнаруженного) [22] для существующей модели.

Количественная оценка предсказательной способности модели на тех же данных, на которых она была разработана (часто называемая предполагаемой эффективностью, *apparent performance*) (см. **рис. 1**), будет давать слишком оптимистичную оценку эффективности модели из-за чрезмерной аппроксимации (переобучения — слишком мало событий исхода, по сравнению с числом потенциальных предикторов, *overfitting*) и использования стратегий выбора предикторов (*predictor selection strategies*) [23–25]. Поэтому исследования, где разрабатывают новые предсказательные модели, всегда должны включать какие-либо формы внутренней проверки (*internal validation*) для количественной оценки любого оптимизма эффективности предсказаний [например, калибровка (*calibration*) и различение (*discrimination*)] и последующей корректировки модели. Внутренняя проверка предполагает использование только исходной выборки (*original study sample*) и включает такие методы, как бутстреппинг (*bootstrapping*) или перекрёстную проверку (*crossvalidation*). Внутренняя проверка — неотъемлемая часть разработки модели [2].

После разработки предсказательной модели настоятельно рекомендуется оценить её эффективность (*performance*) на других данных (не тех, которые использовали для разработки модели). Внешняя проверка (*external validation*) (**вставка В** и **рис. 1**) [20, 26] требует, чтобы для каждого индивидуума в новом наборе данных предсказания исхода были выполнены исходной моделью [т.е. опубликованной моделью *(published model*) или по формуле регрессии (*regression formula*)], а полученные результаты сравнены с фактическими исходами. Внешняя проверка может выполняться с использованием данных, собранных теми же исследователями, обычно с учётом тех же предикторов, определений исхода и способов их регистрации, что и в исходной модели, но в более поздний период — временна́я или ограниченная проверка (*temporal/narrow validation*); или собранных другими исследователями в другом лечебном учреждении или другой стране (что случается достаточно редко [27]) с учётом других определений и способов регистрации — географическая (*geographic validation*) или широкая проверка (*broad validation*); или собранных у схожих участников, но в других условиях [например, модель, разработанная на основе данных учреждений, оказывающих специализированную медицинскую помощь (*secondary care*), оценивается на схожих пациентах из учреждений первичной медицинской помощи (*primary care*)]; либо собранных у участников другого типа [например, модель разрабатывают на взрослых пациентах, а проверяют на детях; или модель, разработанную для прогнозирования фатальных событий (*fatal events*), проверяют на данных о нефатальных событиях (*nonfatal events*)] [19, 20, 26, 28–30]. В случае низкой эффективности (например, при систематических ошибках калибровки), определенной при внешней проверке, модель может быть обновлена или скорректирована (например, путём повторной калибровки или добавления нового предиктора) с использованием проверочного набора данных (*validation data set*) (**вставка В**) [2, 20, 21, 31].

Случайное разделение одного набора данных на две отдельные группы (для разработки и проверки модели) — частое явление в исследованиях предсказательных

моделей. Такой метод ошибочно считают примером внешней проверки. Однако такой подход является слабой и неэффективной формой внутренней проверки, поскольку для разработки модели используются не все имеющиеся данные [23, 32]. Если доступный набор данных достаточно велик, то более эффективным подходом будет его разделение по времени сбора с разработкой модели на данных одного периода и оценкой её эффективности на данных другого периода (временна́я проверка, *temporal validation*). Разделение одного набора данных по временны́м периодам для целей проверки разрабатываемой модели считается промежуточным этапом между её внутренней и внешней проверкой.

Вставка В. Типы исследований предсказательных моделей

Исследования по разработке предсказательной модели без проверки* на независимых данных нацелены на разработку одной (или более) прогностической или диагностической предсказательной модели на основе имеющегося набора данных (*development set*). В таких исследованиях, как правило, определяют значимые для исхода предикторы, каждому предиктору в многомерном анализе присваивают скорректированные коэффициенты, разрабатывают модель для индивидуальных предсказаний, выполняют количественную оценку предсказательной эффективности (*predictive performance*) модели (например, таких её параметров, как различение, калибровка, классификация) на данных, использованных для её разработки. Иногда на этапе разработки модели исследователи могут выполнять количественную оценку дополнительной предсказательной значимости (*incremental / added predictive value*) конкретного предиктора (например, недавно обнаруженного). Результаты исследований, где при разработке модели используют небольшие наборы данных, могут оказаться чересчур оптимистичными. В таких случаях корректность модели проверяют методами многократной/повторной генерации выборок [бутстреппинг (*bootstrapping*), метод складного ножа (*jack-knife*), перекрёстная проверка (*cross-validation*)]. Эти методы позволяют количественно оценить оптимистичность предсказательной эффективности разработанной модели и ожидаемые характеристики модели при её применении у других представителей популяции, из которой был получен набор данных, использованный для разработки модели (*source population*) (см. **рис. 1**). Методы многократной/повторной генерации выборок (*resampling techniques*) часто называют внутренней проверкой модели (*internal validation of the model*), поскольку здесь задействованы лишь те данные, которые использовали для разработки этой модели. Внутренняя проверка — неотъемлемая часть исследований по разработке предсказательных моделей (см. **рис. 1** и **вставку Е**).

Исследования по разработке предсказательных моделей с проверкой* на независимых данных преследуют те же цели, что и предыдущий тип исследований с той лишь разницей, что количественную оценку эффективности модели выполняют на основе данных, которые не использовали в разработке модели (см. **рис. 1**). Такие данные могут быть собраны, например: 1) теми же исследователями с учётом тех же предикторов, определений и оценок исходов, что и в исходной модели, но отобранных из более позднего периода — временна́я или точная проверка (*temporal/narrow validation*); 2) другими исследователями в другом лечебном учреждении или другой стране (что случается достаточно редко [27]) с учётом других определений и оценок — географическая или широкая валидация (*geographic/broad validation*); 3) у участников со схожими характеристиками, но в других условиях [например, модель, изначально разрабатываемая на основе данных пациентов, которые обратились за специализированной медицинской помощью (*secondary care*), затем оценивается на схожих данных пациентов, обратившихся за первичной медицинской помощью (*primary care*)]; 4) либо у участников другого типа [например, модель разрабатывают на взрослых пациентах, а проверяют на детях; или модель, разработанную для прогнозирования фатальных событий (*fatal events*), затем оценивают на основе данных о нефатальных событиях (*nonfatal events*)]. Произвольное разделение данных на две отдельные группы (для разработки и проверки модели) часто ошибочно называют внешней проверкой (*external validation*) модели. В действительности это неэффективная форма внутренней, а не внешней проверки, поскольку различия двух наборов данных обусловлены только случайностью (см. **рис. 1**).

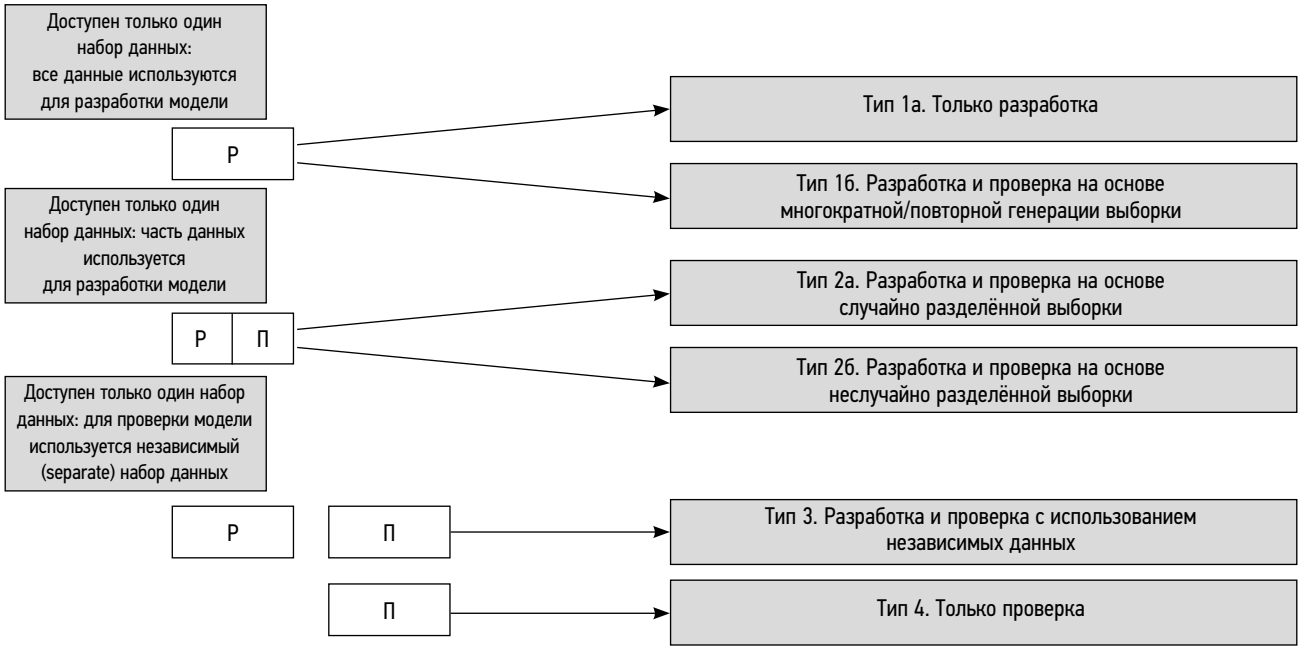
Исследования по проверке* модели с её обновлением или без него посвящены оценке и сравнению предсказательной эффективности одной или нескольких существующих моделей на основе данных, которые не использовали при разработке предсказательной модели. Если модель оказалась неэффективной, за проверочным исследованием может последовать обновление (*updating*) или корректировка модели (*adjusting*) (например, повторная калибровка или расширение модели путём добавления новых предикторов). Теоретически исследование может быть нацелено только на обновление существующей модели с использованием нового набора данных, хотя на практике это маловероятно и даже нежелательно без предварительной проверки оригинальной модели на новых данных (см. **рис. 1**).

* Термин «проверка» (*валидация*) хотя и широко используется, но может вводить в заблуждение, позволяя считать, что исследования по проверке моделей якобы дают ответ «да» (хороший результат) или «нет» (плохой результат) касательно эффективности модели.

Однако цель проверки модели состоит в том, чтобы оценить (количественно) предсказательные характеристики модели либо на основе повторно используемых данных, применённых для разработки модели (так называемая *внутренняя проверка*), либо на других независимых данных, которые не были использованы на этапе разработки модели (*внешняя проверка*).

НЕПОЛНЫЕ И НЕТОЧНЫЕ ОТЧЁТЫ

В медицинской литературе появляется всё больше публикаций, посвящённых предсказательным моделям [9, 33, 34]. Более того, лица, определяющие политику в области здравоохранения, всё чаще предлагают их использовать в клинических рекомендациях [35–40]. Для некоторых заболеваний существует большое количество конкурирующих предсказательных моделей для одних и тех же исходов или целевой популяции. Например, описано более 100 прогностических моделей для предсказания исхода после травм головного мозга [41], более 100 моделей для рака предстательной железы [42], более 60 моделей для рака молочной железы [43], 45 моделей для прогноза развития сердечно-сосудистых событий после диагностики сахарного диабета [44], более 40 моделей для предсказания случаев сахарного диабета



Тип анализа	Описание
Тип 1а	Разработка предсказательной модели, характеристики которой затем прямо оцениваются на точно таких же данных (предполагаемая эффективность, *apparent performance*).
Тип 1б	Разработка предсказательной модели с использованием всего набора данных, но где последующая оценка эффективности и уверенности в успешности разработанной модели выполняется на основе многократной/повторной генерации выборки (например, бутстреппинг или перекрёстная проверка). Эти методы, обычно относимые к внутренней проверке, рекомендуются в качестве необходимого условия для разработки предсказательной модели, особенно при ограниченном количестве данных [6, 14, 15].
Тип 2а	Данные в случайном порядке разделяют на две группы: для разработки модели и для оценки её качественных (предсказательных) характеристик. Этот дизайн обычно не рекомендуется, или он не лучше, чем тип 1б, особенно при ограниченном количестве данных, поскольку приводит к потере мощности при разработке и проверке модели [14, 15, 16].
Тип 2б	Данные в неслучайном порядке (например, по местоположению или времени) разделяют на две группы: для разработки модели и для оценки её качественных (предсказательных) характеристик. Этот тип исследования более предпочтителен для оценки модели, чем предыдущий, поскольку допускает неслучайные вариации между двумя наборами данных [6, 13, 17].
Тип 3	Разработка предсказательной модели с использованием одного набора данных и оценка её качественных (предсказательных) характеристик на основе независимых данных (например, из другого исследования).
Тип 4	Оценка качественных (предсказательных) характеристик существующей (опубликованной) предсказательной модели на основе независимых данных [13].
Типы 3 и 4 обычно относят к исследованиям внешней проверки моделей. Возможно, таковым является тип 2б, хотя его можно рассматривать и как промежуточный этап между внутренней и внешней проверкой	

Примечание. Р — разработка; П — проверка.

Рис. 1. Типы исследований предсказательных моделей, охватываемые рекомендациями TRIPOD.

2-го типа [45] и 20 моделей для оценки риска продолжительного пребывания в отделении интенсивной терапии после кардиохирургических вмешательств [46].

Учитывая обилие опубликованных научных работ о предсказательных моделях практически во всех областях медицины, критическая оценка и обобщение представленных отчётов — основное требование, позволяющее читателям, поставщикам медицинских услуг и лицам, принимающим решения в сфере здравоохранения, судить, какие модели в каких ситуациях могут быть полезны. Такая оценка, в свою очередь, возможна, если в рукописях чётко и ясно описаны ключевые детали разработки и проверки предсказательных моделей [47, 48]. Только так можно объективно оценить обобщаемость (*generalizability*) и риск систематических ошибок (*risk of bias*) опубликованной предсказательной модели [49, 50], а другим исследователям при необходимости воспроизвести полученные результаты на основе тех же данных [51, 52]. Однако многие обзоры показали, что качество отчётов, в которых описана разработка и проверка предсказательных моделей для различных заболеваний, опубликованных в разных журналах, является низким [3, 34, 41, 43, 45, 46, 48, 53–95]. Например, в обзоре новых предсказательных моделей по онкологическим заболеваниям было показано, что отчёты были крайне низкого качества, а все аспекты разработки моделей скудно освещены. Такой же вывод сделан при изучении обзора предсказательных моделей для случаев сахарного диабета 2-го типа [45] и предсказательных моделей, опубликованных в 6 влиятельных общемедицинских журналах [34].

Руководства по представлению результатов рандомизированных исследований (CONSORT [96]), наблюдательных исследований (STROBE [97]), исследований маркеров опухолей (REMARK [98]), молекулярной эпидемиологии (STROBE-ME [99]), диагностических исследований (STARD [100]) и исследований предсказания генетического риска (GRIPS [101]) также содержат пункты, относящиеся ко всем типам исследований, включая исследования, в которых разрабатывались или проверялись предсказательные модели. Из выше перечисленных лишь два руководства наиболее тесно связаны с предсказательными моделями — REMARK и GRIPS. Однако в контрольном перечне рекомендаций REMARK основное внимание уделяется прогностическим факторам, а не предсказательным моделям, в то время как рекомендации GRIPS направлены на повышение качества отчётов, содержащих описание предсказаний на основе генетических факторов риска и специфических методологических вопросов, связанных с обработкой данных большого количества генетических вариантов.

Рекомендации TRIPOD были разработаны для более широкого круга исследований. Они актуальны для разработки и проверки предсказательных моделей как для диагностики, так и для прогнозирования во всех областях медицины и для всех типов предикторов. TRIPOD уделяет значительное внимание исследованиям по проверке моделей и требованиям к оформлению отчётов о таких исследованиях.

РЕКОМЕНДАЦИИ TRIPOD

Исследования предсказательных моделей можно разделить на 5 обширных категорий [1, 8–10, 19, 20, 28, 33, 102–104]: 1) исследования прогностических или диагностических предикторов; 2) исследования по разработке предсказательных моделей без внешней проверки; 3) исследования по разработке предсказательных моделей с внешней проверкой; 4) проверочные исследования предсказательных моделей; 5) исследования влияния предсказательных моделей. Рекомендации TRIPOD касаются исследований, целью которых является разработка или проверка одной или нескольких предсказательных моделей (**вставка B**). Эти исследования, в свою очередь, можно также разделить на несколько типов (**рис. 1**). Всё большее количество исследований оценивают добавленную прогностическую значимость (*incremental value*) [103] конкретного предиктора, что позволяет понять, нуждается ли модель в обновлении или корректировке [22, 105, 106]. Такие исследования также охватываются рекомендациями TRIPOD (**вставка B** и **рис. 1**).

Исследования прогностических или диагностических предикторов и исследования влияния предсказательных моделей часто имеют разные цели, дизайн и особенности представления полученных результатов, по сравнению с исследованиями, посвящёнными разработке или проверке предсказательных моделей. Первые обычно направлены на выявление предикторов, независимо (от других известных предикторов) ассоциированных с конкретным прогностическим или диагностическим результатом. Они не нацелены на разработку окончательной предсказательной модели, которую будут впоследствии использовать для индивидуальных прогнозов в отношении других пациентов. Исследования влияния модели направлены на количественную оценку эффекта (влияния, *impact*) использования модели на принятие решений пациентом и врачом или непосредственно на оценку важных для здоровья исходов, по сравнению с её неиспользованием [20, 102, 107]. Такие исследования следуют плану сравнительного интервенционного исследования, а не плану простого (с одной выборкой) когортного исследования, используемого при разработке и проверке моделей, и в идеале должны быть спланированы как (кластерные) рандомизированные исследования. И хотя многие пункты настоящего публикационного руководства применимы и к этих двум типам предсказательных исследований, другие рекомендации по отчётности могут быть более подходящими. Так, рекомендации REMARK разработаны для повышения качества отчётов о прогностических исследованиях (как правило, одного фактора) [98, 108], а стандарты CONSORT [96, 109]

и STROBE [97] — рандомизированных и нерандомизированных исследований влияния предсказательных моделей.

Кроме того, TRIPOD в первую очередь предназначен для описания исследований предиктивных моделей для исходов бинарного вида (*binary outcomes*) (заболевание присутствует или отсутствует) или исходов, содержащих информацию о времени до наступления события (*time-to-event outcomes*) (например, 10-летний риск развития заболеваний сердечно-сосудистой системы), поскольку это наиболее распространённые типы исходов, которые можно предсказывать в медицине. Однако исходы могут быть представлены и в виде непрерывных (*continuous measurements*) (например, артериальное давление, размер опухоли, процент стеноза сосудов, уровень интеллекта, качество жизни или продолжительность госпитализации), номинальных (*nominal outcomes*) (например, разные диагнозы, а не простая констатация наличия или отсутствия целевого заболевания; тип инфекции, определяемой как вирусная, бактериальная, или её отсутствие) и ординальных переменных (*ordinal outcomes*) (например, стадия рака, оценка по шкале комы Глазго [110], шкалы Ранкина [111]), для предсказания которых могут также разрабатывать модели [2, 112]. Большинство рекомендаций и пунктов отчётности TRIPOD в равной степени применимы к описанию исследований разработки или проверки предсказательных моделей для таких исходов.

Более того, TRIPOD сфокусирован на предсказательных моделях, разработанных с помощью метода регрессионного анализа, поскольку с помощью этого подхода разрабатывают, проверяют или обновляют большинство таких моделей в медицинских исследованиях. Однако большинство пунктов руководства в равной степени применимы к предсказательным моделям, разработанным, проверенным или обновлённым с помощью других методов (например, деревья классификации, нейронные сети, генетическое программирование, алгоритм «случайный лес», векторные методы машинного обучения). Основное отличие этих подходов от регрессионного моделирования заключается лишь в методе анализа данных для получения предсказательной модели. Вместе с тем проблемы прозрачности описания этих нерегрессионных подходов к моделированию вызывают особую озабоченность, особенно в отношении последующей воспроизводимости результатов исследования и их внедрения в практику.

К ИСТОРИИ РАЗРАБОТКИ РЕКОМЕНДАЦИЙ TRIPOD

Мы следовали опубликованному руководству по разработке рекомендаций по оформлению отчётов об исследованиях [113] и создали координационный комитет (д-ра Collins, Altman, Moons, Reitsma) для организации и координации разработки TRIPOD. Мы провели систематический поиск в MEDLINE, EMBASE, PsycINFO и Web of Science для обнаружения любых опубликованных статей с рекомендациями по составлению отчётов о многофакторных предсказательных моделях или методологических аспектах их разработки или проверки, обзоров опубликованных отчётов о многофакторных предсказательных моделях, в которых оценивались методология или качество отчётов, а также обзоров методологии и качества отчётов о многофакторных моделях в целом. В результате составили предварительный контрольный перечень (*cheklist*) из 129 пунктов, который затем сократили до 76 пунктов для обсуждения в экспертном сообществе.

Для участия в онлайн-опросе и оценки важности 76 пунктов указанного перечня по электронной почте были приглашены 25 экспертов, специализирующихся на предсказательных моделях. В число участников (24 из 27 человек) входили методологи, медицинские работники и редакторы журналов (помимо 25 приглашенных экспертов в опросе участвовали два статистических редактора журнала *Annals of Internal Medicine).*

24 эксперта (22 из которых участвовали в опросе) присутствовали на трехдневной встрече в Оксфорде (Великобритания) в июне 2011 г. Эта междисциплинарная группа включала статистиков, эпидемиологов, методологов, медицинских работников и редакторов журналов (**Приложение**) [114]. Некоторые из членов группы уже имели опыт разработки руководств по отчётности для других видов клинических исследований.

В ходе встречи эксперты проанализировали результаты опроса и обсудили каждый из 76 пунктов-кандидатов контрольного перечня. По каждому пункту был достигнут консенсус относительно того, следует ли его сохранить, объединить с другим или исключить из перечня. Участников встречи также просили предлагать дополнительные пункты. Затем контрольный перечень был пересмотрен членами координационного комитета в ходе многочисленных личных встреч и вновь разослан участникам для окончательного утверждения. При внесении изменений были предприняты сознательные усилия по согласованию наших рекомендаций с другими руководствами, и, где это возможно, мы выбрали ту же или похожую формулировку для пунктов контрольного перечня.

РЕКОМЕНДАЦИИ TRIPOD: РАЗЪЯСНЕНИЯ И УТОЧНЕНИЯ

Цель и структура документа

TRIPOD — это контрольный перечень из 22 пунктов с рекомендациями по оформлению отчётов исследований, посвящённых разработке или проверке многофакторных предсказательных моделей (**табл. 1**) [114]. Предложенные рекомендации касаются таких разделов научных рукописей, как название и аннотация (пункты 1, 2), обоснование и цели исследования (пункт 3), методы

Таблица 1. Контрольный перечень пунктов для включения в отчёты об исследованиях по разработке или проверке многофакторных предсказательных моделей для диагностики или прогноза*

Раздел/тема	Пункт	Разработка или проверка?	Пункты контрольного перечня	Стр.
Название и резюме				
Название	1	Р, П	Обозначьте цель исследования [разработка и (или) проверка многофакторной предсказательной модели], целевую популяцию и предсказываемый исход	
Резюме	2	Р, П	Представьте краткое описание целей, дизайна исследования, условий его проведения, участников, размера выборки, предикторов, исход, статистического анализа, результатов и выводов	
Введение				
Обоснование и цели	3а	Р, П	Обозначьте медицинский контекст темы исследования (в том числе диагностический или прогностический), обоснуйте необходимость разработки или проверки многофакторной предсказательной модели, приведите ссылки на существующие модели	
	3б	Р, П	Укажите цели исследования, упомянув, идет ли речь о разработке и (или) проверке модели	
Методы				
Источник данных	4а	Р, П	Опишите дизайн исследования или источник данных (например, данные рандомизированного или когортного исследования, регистра), отдельно для наборов данных, использованных для разработки и проверки модели, если применимо	
	4б	Р, П	Укажите основные даты исследования, включая начало и завершение набора участников и, если применимо, завершение периода последующего наблюдения	
Участники	5а	Р, П	Опишите условия и место проведения исследования (например, учреждения первичной или специализированной медицинской помощи, общая популяция), указав количество и местонахождение участвующих центров	
	5б	Р, П	Опишите критерии отбора участников	
	5в	Р, П	Подробно опишите медицинское вмешательство, если применимо	
Исход	6а	Р, П	Определите предсказываемый моделью исход, включив описание способов и сроков его регистрации	
	6б	Р, П	Сообщите о любых действиях для маскирования (ослепления) при оценке предсказываемого исхода	
Предикторы	7а	Р, П	Опишите все предикторы, использованные при разработке многофакторной предсказательной модели, указав, как и когда они были измерены	
	7б	Р, П	Сообщите о действиях для маскирования (ослепления) при оценке предикторов исхода или любых других предикторов	
Размер выборки	8	Р, П	Объясните, как был определён размер выборки исследования	
Отсутствующие данные	9	Р, П	Опишите, как обрабатывали отсутствующие (неполные) данные (например, анализ только полных наблюдений, подстановка значений), детально – применение любого метода подстановки значений	
Методы статистического анализа	10а	Р	Опишите, как поступали с предикторами в процессе анализа данных	
	10б	Р	Укажите тип модели, последовательность её построения (включая выбор предикторов) и методы внутренней проверки	
	10в	П	Для проверочных исследований – опишите, как рассчитывали вероятности предсказываемого исхода	
	10г	Р, П	Укажите все показатели, с помощью которых оценивали эффективность модели и, если применимо, сравнивали несколько моделей	
	10д	П	Опишите любое обновление модели (например, повторную калибровку), выполненное в результате её проверки (если применимо)	

Таблица 1. Окончание

Раздел/тема	Пункт	Разработка или проверка?	Пункты контрольного перечня	Стр.
Группы риска	11	Р, П	Подробно опишите, как определяли группы риска (если применимо)	
Разработка против проверки	12	П	В проверочном исследовании укажите любые отличия в условиях проведения, критериях отбора, исходе и предикторах от таковых в исследовании, в котором модель была разработана	
Результаты				
Участники	13а	Р, П	Опишите поток участников в ходе исследования, включая количество участников с исходом и без него, и, если применимо, характеристики периода отслеживания исходов. Графическое представление этой информации может быть полезным	
	13б	Р, П	Опишите характеристики участников исследования (основные демографические и клинические показатели, доступные предикторы), укажите количество участников с отсутствующими данными по показателям предикторов и исхода	
	13в	П	Для проверочных исследований – представьте сравнение распределения важных переменных (демографические показатели, предикторы, исход) с данными, использованными для разработки модели	
Разработка модели	14а	Р	Укажите количество участников и событий исхода для каждого анализа	
	14б	Р	Если применимо, укажите нескорректированные оценки ассоциации каждого потенциального предиктора и исхода	
Характеристики модели	15а	Р	Представьте полную предсказательную модель, позволяющую предсказывать исход для отдельных лиц (т.е. все коэффициенты регрессии и свободный коэффициент модели или исходный показатель выживаемости в определённый момент времени)	
	15б	Р	Объясните, как использовать предсказательную модель	
Эффективность модели	16	Р, П	Сообщите показатели эффективности (включая доверительные интервалы) предсказательной модели	
Обновление модели	17	П	Если применимо, сообщите результаты любого обновления модели (т.е. состава модели, условий её применения, характеристик эффективности)	
Обсуждение				
Ограничения	18	Р, П	Обсудите любые ограничения исследования (например, нерепрезентативная выборка, недостаточное количество событий на один предиктор, отсутствующие данные)	
Интерпретация результатов	19а	П	В случае проверочного исследования обсудите полученные результаты с упоминанием характеристик оригинальной модели, а также характеристик, полученных с использованием любых других проверочных данных	
	19б	Р, П	Обсудите результаты с учётом целей, ограничений, результатов схожих исследований и других актуальных сведений	
Применение	20	Р, П	Обсудите потенциал клинического использования модели и значение для будущих исследований	
Другие сведения				
Дополнительная информация	21	Р, П	Предоставьте информацию о доступности дополнительных материалов, таких как протокол исследования, веб-калькулятор и наборы данных	
Финансирование	22	Р, П	Укажите источник финансирования и роль спонсоров в настоящем исследовании	

* Р — пункты, относящиеся только к разработке предсказательной модели; П — пункты, относящиеся только к проверке предсказательной модели; Р, П — пункты, относящиеся и к разработке, и к проверке предсказательной модели. Рекомендуем использовать контрольный перечень TRIPOD в сочетании с разъяснениями и уточнениями, изложенными в настоящей статье.

(пункты 4–12), результаты (пункты 13–17), обсуждение (пункты 18–20) и дополнительная информация (пункты 21, 22). Рекомендации TRIPOD применимы к исследованиям, посвящённым исключительно разработке, разработке и внешней проверке или исключительно внешней проверке (с последующим обновлением или без него) диагностической или прогностической предсказательной модели (**вставка В**). Поэтому некоторые пункты (обозначенные буквой Р) касаются только разработки предсказательной модели (пункты 10а, 10б, 14, 15), другие (обозначенные буквой П) — только проверки модели (пункты 10в, 10д, 12, 13в, 17, 19а). Остальные пункты применимы для описания всех типов исследований разработки и проверки предсказательных моделей (Р, П).

Обсуждение и объяснение всех пунктов контрольного перечня TRIPOD представлены в **табл. 1**. Для большей ясности мы разделили обсуждение сложных и объёмных пунктов на несколько подпунктов.

Цель этого разъясняющего и уточняющего документа — обозначить структуру отчётов об исследованиях предсказательных моделей. Многие исследования такого рода методологически слабы, поэтому в этом документе мы суммируем характеристики хороших (и ограничения менее убедительных) исследований независимо от того, как они представлены.

Использование примеров

По каждому пункту мы приводим примеры из опубликованных статей с результатами как разработки, так и проверки предсказательных моделей, часто моделей для диагностики и прогноза; они иллюстрируют ту информацию, о которой следует сообщать. Это не означает, что исследование, из которого был заимствован пример, было качественно выполнено и представлено, или описанные методы обязательно являются лучшим решением для исследований предсказательных моделей. Наш выбор примеров скорее обусловлен корректной иллюстрацией конкретного аспекта того или иного пункта, правильно представленного в контексте методов, использованных авторами исследования. Некоторые примеры были отредактированы (сокращён текст, добавлены примечания и аббревиатуры, удалены цитаты, упрощены таблицы).

Использование TRIPOD

В зависимости от типа исследования предсказательной модели [разработка и (или) проверка] каждый пункт контрольного перечня (подходящий типу исследования) должен быть рассмотрен в представленном отчёте. Если какой-то из пунктов не может быть включён в отчёт, следует сообщить об отсутствии информации по данному пункту или его неприменимости. Многие пункты упорядочены естественным (привычным) образом, но не все. Мы не настаиваем на определённом порядке изложения информации, поскольку он может зависеть от политики форматирования журнала. Например, авторы могут сообщать некоторые данные в дополнительных разделах, таких как онлайн-приложения.

Чтобы облегчить работу редакторов, рецензентов и в конечном счёте читателей, мы рекомендуем присылать контрольный перечень дополнительным файлом с указанием страниц, на которых представлена информация по каждому пункту. Шаблон отчёта TRIPOD доступен по адресу www.tripod-statement.org.

С анонсами и дополнительной информацией о рекомендациях TRIPOD можно ознакомиться на нашей странице в Twitter (@TRIPODStatement). В целях распространения и продвижения рекомендаций TRIPOD документ размещён в библиотеке проекта EQUATOR Network (Повышение качества и прозрачности исследований в области здравоохранения) (http://www.equator-network.org).

КОНТРОЛЬНЫЙ ПЕРЕЧЕНЬ TRIPOD

Название и резюме

Название

Пункт 1. Обозначьте цель исследования [разработка и (или) проверка многофакторной предсказательной модели], целевую популяцию и предсказываемый исход [Р; П[1]].

Примеры

«Разработка и проверка клинической шкалы для определения вероятности поражения коронарных артерий у мужчин и женщин с подозрением на ишемическую болезнь сердца» [115]. (Диагностика; Разработка; Проверка.)

«Разработка и проверка на внешних данных модели для прогнозирования двухлетней выживаемости пациентов с немелкоклеточным раком лёгких после химиолучевой терапии» [116]. (Прогнозирование; Разработка; Проверка.)

«Предсказание 10-летнего риска развития сердечно-сосудистого заболевания в Великобритании: независимая проверка на внешних данных обновлённой версии QRISK2» [117]. (Прогнозирование; Проверка.)

«Разработка модели для предсказания 10-летнего риска развития гепатоцеллюлярной карциномы у японцев в среднем возрасте: второе проспективное когортное исследование Центра общественного здравоохранения Японии» [118]. (Прогнозирование; Разработка.)

Пример с дополнительной информацией

«Разработка и проверка алгоритма на основе логистической регрессии для оценки вероятности обнаружения ишемической болезни сердца до и после нагрузочного теста» [119]. (Диагностика; Разработка; Проверка.)

[1] Здесь и далее для каждого пункта контрольного перечня заглавной буквой указано назначение исследования, где Р — разработка, П — проверка.

Примеры известных моделей

«Проверка шкалы Framingham для оценки риска развития ишемической болезни сердца: результаты исследования в нескольких этнических группах» [120]. (Прогнозирование; Проверка.)

«Проверка прогностических моделей SAPS II, APACHE II и APACHE III на внешних данных, полученных в Южной Англии: многоцентровое исследование» [121]. (Прогнозирование; Проверка.)

Пояснение

Название статьи должно быть информативным, но при этом не слишком длинным, чтобы облегчить поиск исследований потенциальными читателями или исследователями, проводящими систематические исследования многофакторных предсказательных моделей. В идеале авторы могут указать в заголовке 4 основных элемента: 1) тип исследования (разработка, проверка модели или и то, и другое); 2) клинический контекст (диагностика или прогнозирование); 3) целевая популяция (лица или пациенты, для которых предназначена модель); 4) исход, предсказываемый моделью.

Исследования предсказательных моделей посвящены разработке моделей (включая внутреннюю проверку; пункт 10б) и (или) их внешней проверке (**вставка B** и **рис. 1**). Авторы должны явно указывать тип своего исследования, используя эти термины в заголовке. Если цель исследования — обновление ранее разработанной модели или оценка дополнительной ценности определённого предиктора, следует сообщить об этом. Более того, поскольку многие читатели заинтересованы в поиске доступной литературы о конкретной популяции или субпопуляциях, пациентах или о конкретном исходе у этих лиц, такие идентифицирующие термины полезно включить в заголовок.

Как видно из выше приведённых примеров, все эти аспекты можно обозначить в названии рукописи, не создавая длинные заголовки. Авторы, проводившие проверку модели на внешних данных в качестве единственной цели или в связи с разработкой предсказательной модели, должны указывать это в названии.

Термины «диагностический» (*diagnostic*) или «прогностический» (*prognostic*) не часто приводятся в заголовке статей, но об этом может косвенно свидетельствовать описание исследуемой популяции или исходов. Например, названия, включающее «...у мужчин и женщин с подозрением на ишемическую болезнь сердца» явно указывает на то, что это исследование диагностической модели [115]. Названия некоторых предсказательных моделей настолько хорошо известны, что заголовки последующих проверочных исследований не содержат указания на целевую популяцию или предсказываемый исход. Однако если исследование нацелено на проверку известной модели в других условиях или предсказывание другого исхода, это должно быть чётко указано в заголовке.

В некоторых случаях для уточнения характера исследования в заголовке статьи можно указать тип предикторов (например, предикторы из анамнеза пациентов или их осмотра), сроки предсказания (например, предсказание послеоперационных исходов на основе дооперационных характеристик пациента) или сроки наступления исхода (например, 10-летний риск развития сердечно-сосудистых заболеваний), но без его чрезмерного увеличения.

В недавнем обзоре 78 проверочных исследований, в которых были использованы внешние данные, в заголовке лишь 21 статьи (27%) присутствовали термины «валидация» (*validation*) или «валидность» (*validity*). И только в заголовке одной статьи авторы явным образом указали, что проверка была выполнена независимыми исследователями [122].

Резюме

Пункт 2. Представьте краткое описание целей, дизайна исследования, условий его проведения, участников, размера выборки, предикторов, исхода, статистического анализа, результатов и выводов (Р; П).

Примеры

«ЦЕЛЬ. Разработать и проверить модель для прогнозирования ранней смерти пациентов с кровотечением, вызванным травмой.

ДИЗАЙН. Многофакторная логистическая регрессия большой международной когорты пациентов с травмами.

УСЛОВИЯ ПРОВЕДЕНИЯ. 274 больницы в 40 странах с высоким, средним и низким уровнем дохода.

УЧАСТНИКИ. Разработка прогностической модели: 20 127 травмированных пациентов с сильным кровотечением или риском его возникновения в течение 8 ч после травмы, которые приняли участие в испытании Clinical Randomisation of an Antifibrinolytic in Significant Haemorrhage (CRASH 2). Внешняя проверка: 14 220 пациентов с травмами, отобранных из базы данных Trauma Audit and Research Network (TARN), которая включает в основном пациентов из Великобритании.

ИСХОДЫ. Смерть в больнице в течение 4 недель после травмы.

РЕЗУЛЬТАТЫ. 3076 человек (15%) умерли в исследовании CRASH 2 и 1765 (12%) по данным TARN. Оценка по шкале комы Глазго, возраст и систолическое артериальное давление были наиболее сильными предикторами наступления летального исхода. Другими предикторами, включёнными в окончательную модель, были географический регион (страна с низким, средним или высоким уровнем дохода), частота сердечных сокращений, время после травмы и её тип. Дискриминация и калибровка были удовлетворительными (*c*-индекс выше 0,80 как в CRASH 2, так и TARN). Построили простую диаграмму для определения вероятности наступления смерти в месте оказания медицинской помощи. Для более подробной оценки риска доступен веб-калькулятор (http: //crash2.lshtm.ac.uk).

ВЫВОДЫ. Прогностическая модель может быть использована для получения обоснованных предсказаний наступления смерти у пациентов с травматическим кровотечением и сортировки пациентов и потенциально может помочь сократить время до диагностических и спасающих жизнь процедур (визуальная диагностика, хирургическое вмешательство, введение транексамовой кислоты). Будучи важным прогностическим фактором, возраст особенно актуален для стран с высоким уровнем доходов и большим количеством травмированных пациентов пожилого возраста» [123]. (Прогнозирование; Разработка.)

«ЦЕЛЬ. Проверить и уточнить ранее разработанные правила принятия клинических решений, которые помогают эффективно использовать рентгенографию при острых травмах голеностопного сустава.

ДИЗАЙН. Исследование проводили в два этапа: проверка и уточнение оригинальных правил (первый этап) и проверка уточнённых правил (второй этап).

УСЛОВИЯ ПРОВЕДЕНИЯ. Отделения неотложной медицинской помощи двух университетских больниц.

ПАЦИЕНТЫ. Удобная выборка (*convenience sample*) взрослых с острыми травмами голеностопного сустава (1032 из 1130 подходящих пациентов на первом этапе и 453 из 530 — на втором).

ОСНОВНЫЕ ПОКАЗАТЕЛИ ИСХОДОВ. Врачи отделения неотложной медицинской помощи оценивали каждого пациента по стандартизированным клиническим параметрам и классифицировали потребность в рентгенографии в соответствии с исходными (первый этап) и уточнёнными (второй этап) правилами принятия решения. Правила принятия решений оценивали по их способности корректно определить стандартный критерий переломов на рентгенограммах голеностопного сустава и стопы. Оригинальные правила принятия решений были уточнены методами однофакторного анализа и рекурсивного разделения (*recursive partitioning*).

ОСНОВНЫЕ РЕЗУЛЬТАТЫ. На первом этапе установлено, что оригинальные правила принятия решений имели чувствительность 1,0 [95% доверительный интервал (ДИ) 0,97–1,0] при диагностике 121 перелома мыщелков большеберцовой кости и 0,98 (95% ДИ 0,88–1,0) — 49 переломов среднего отдела стопы. У 116 пациентов коэффициент каппа составлял 0,56 для рентгенограмм голеностопного сустава и 0,69 — стопы. Рекурсивное разделение 20 предикторных переменных позволило уточнить правила принятия решений для рентгенограмм голеностопного сустава и стопы. На втором этапе установлено, что уточнённые правила принятия решений имели чувствительность 1,0 (95% ДИ 0,93–1,0) при диагностике 50 переломов мыщелков большеберцовой кости и 1,0 (95% ДИ 0,83–1,0) — 19 переломов среднего отдела стопы. Потенциальное сокращение (оценка) количества рентгенографических обследований для выявления переломов голеностопа составило 34%, переломов стопы — 30%. Вероятность перелома, если соответствующее правило принятия решения было отрицательным, оценивается в 0% (95% ДИ 0–0,8) для голеностопных суставов и 0% (95% ДИ 0–0,4%) для стопы.

ЗАКЛЮЧЕНИЕ. В результате уточнения и проверки установлено, что правила Ottawa для голеностопных суставов на 100% чувствительны к переломам, надёжны и потенциально могут позволить врачам безопасно сократить количество рентгенографических обследований пациентов с травмами голеностопного сустава на одну треть. Полевые испытания позволят оценить возможность внедрения этих правил в клиническую практику» [124]. (Диагностика; Проверка; Обновление.)

Пояснение

Резюме (*abstracts*) содержат основную информацию, которая позволяет читателям оценить методологию и актуальность (*relevance*) исследования, а также ознакомиться с результатами. Резюме может оказаться единственным, что будет легко доступно, и поможет таким образом читателям решить, читать ли полный отчёт. Мы рекомендуем включить в резюме как минимум цели исследования [в идеале с кратким изложением предпосылок (*background*) или обоснования (*rationale*)], описание условий проведения, участников, размера выборки (и число событий), исхода, предикторов, методов статистического анализа, результатов (например, показатели эффективности модели и коэффициенты регрессии), заключение. Структурированное резюме предпочтительнее, хотя требования к их оформлению в разных журналах отличаются.

В резюме должны быть указаны те же атрибуты, что и в заголовке (пункт 1), включая описание цели исследования (разработка или проверка модели, или и то, и другое), типа модели (диагностическая или прогностическая), целевой популяции и предсказываемого исхода. В случае исследований по разработке моделей указание всех потенциальных предикторов может оказаться неосуществимым в силу их большого количества. В этих случаях достаточно назвать их общее количество и основные категории с указанием периода определения (например, при сборе анамнеза и медицинском осмотре). В идеале при описании результатов необходимо указать предикторы, включённые в окончательную модель, а также показатели предсказательной эффективности модели. Это может быть необязательным в случае сложных моделей со множеством предикторов или в исследованиях, в которых проводилась проверка ранее разработанной модели на новых данных.

Информативные резюме и заголовки отчётов об исследованиях предсказательных моделей позволяют исследователям находить подходящие исследования при проведении поиска литературы. Опубликовано несколько стратегий поиска для обнаружения клинических предсказательных моделей [125–127]. Недавно они были протестированы и слегка изменены независимыми исследователями, которые пришли к выводу, что они пропускают

некоторое количество исследований клинических предсказательных моделей (хотя они менее эффективны в поиске других типов исследований по вопросам предсказания) [128]. Также были разработаны специальные поисковые фильтры для обнаружения исследований предсказательных моделей в области первичной медицинской помощи [129].

Введение

Обоснование и цели

Пункт 3a. Обозначьте медицинский контекст темы исследования (в том числе диагностический или прогностический), обоснуйте необходимость разработки или проверки многофакторной предсказательной модели, приведите ссылки на существующие модели (Р; П).

Примеры

«Столкнувшись с острым инфекционным конъюнктивитом, большинство врачей общей практики не могут отличить бактериальную причину болезни от вирусной. На практике более 80% таких пациентов получают антибиотики. А значит, при остром инфекционном конъюнктивите назначают множество ненужных глазных антибиотиков. <...>Чтобы выбрать тех пациентов, которым лечение антибиотиками может принести наибольшую пользу, врачу общей практики необходим информативный диагностический инструмент для определения наличия бактериальной инфекции. С таким инструментом можно сократить количество назначений антибиотиков, а их применение сделать оправданным. Большинство врачей общей практики отличают бактериальную инфекцию от другой причины на основе признаков и симптомов заболевания. Дополнительные диагностические обследования (например, посев отделяемого из конъюнктивы) проводятся редко главным образом из-за длительности таких процедур. Могут ли врачи общей практики дифференцировать бактериальный и вирусный конъюнктивит только на основании признаков и симптомов? <...> Недавно опубликованный систематический обзор не нашёл доказательств этим утверждениям. В настоящей статье представлено первое эмпирическое исследование диагностической информативности признаков и симптомов острого инфекционного конъюнктивита» [130]. (Диагностика; Разработка.)

«В поисках практической прогностической системы для пациентов с карциномой околоушной железы ранее мы создали прогностический индекс на основе анализа пропорциональных рисков Кокса, в исходной популяции из 151 пациента с таким диагнозом в Институте рака Нидерландов. В таблице <...> показаны значения прогностического индекса PS1 до лечения, который объединяет информацию, доступную до операции, и прогностического индекса PS2 после лечения, который включает информацию из операционного образца. Для каждого пациента индекс суммирует надлежащим образом взвешенные важные клинико-патологические характеристики в число, соответствующее предполагаемой вероятности рецидива опухоли. Эти индексы показали хорошую дискриминацию в исходной популяции и в независимой общенациональной базе данных голландских пациентов с карциномой околоушной железы. Согласно Justice и соавт. следующий уровень проверки должен быть международным. <...> С этой целью была создана международная база данных пациентов, получавших лечение в Лёвене и Брюсселе (Бельгия) и в Кёльне (Германия), получены данные о прогностических переменных, необходимых для расчёта индексов, и проведено сравнение предсказаний с исходами. Таким путём мы попытались добиться очередного клинического и статистического подтверждения» [131]. (Прогнозирование; Проверка.)

«Любые пересмотры и обновление модели предсказания риска должны подвергаться постоянной оценке (проверке), чтобы показать, что её польза для повседневной клинической практики осталась прежней, или что её эффективность стала выше благодаря внесённым в модель уточнениям. Мы описываем результаты независимой оценки эффективности QRISK2 2011 на большом наборе данных документации общей врачебной практики в Великобритании, сравнивая её эффективность с более ранними версиями QRISK и скорректированной NICE-версией модели предсказания риска Framingham» [117]. (Прогнозирование; Проверка.)

Пояснение

Многофакторные предсказательные модели могут иметь разное предназначение, поэтому читатели нуждаются в ясном и однозначном описании обоснования необходимости разработки и (или) проверки модели и её потенциального применения. Авторы должны описать конкретный клинический контекст (такой как клиническое решение), в котором модель будет использоваться. Например, диагностическая предсказательная модель может использоваться для принятия решения о назначении инвазивных или более дорогостоящих тестов у определённых пациентов, а прогностическая модель может информировать пациентов с определённым заболеванием о возможном исходе или помочь оценить возможности последующего лечения.

Медицинский контекст и планируемое использование модели обеспечивают обоснование выбора пациентов (включая условия их наблюдения), на кого могут быть распространены результаты предсказания и какие типы предикторов будут доступны в этих условиях и, следовательно, могут быть рассмотрены на предмет включения в модель. Выбор исхода является критическим фактором, определяющим клиническую значимость (*clinical relevance*) модели, поэтому авторам необходимо представить обоснование выбора конкретного исхода. Желательно, чтобы исходы и продолжительность их отслеживания были значимыми для пациентов и принятия клинических решений.

Проблемы могут возникнуть, если используются чрезмерно широкие определения исходов, что повышает вероятность отнесения слишком большого количества лиц

к группе высокого риска [132]. Аналогичная проблема возникает и в диагностике, если новое определение болезни включает отклонения от нормы значений нового чувствительного маркера или на изображениях высокого разрешения, что может привести к гипердиагностике и избыточному лечению [132, 133]. Описание медицинского контекста должно также указывать на любые клинические решения, которые могут быть основаны на прогнозируемом риске. Ниже приведено несколько примеров использования многофакторных предсказательных моделей с диагностической и прогностической целяью.

Возможные варианты клинического использования многофакторных диагностических моделей:

1. Решения о целесообразности назначения инвазивных и дорогостоящих диагностических тестов или направлении пациентов для оказания специализированной медицинской помощи (*secondary care*). Пример: правило *Ottawa* для назначения рентгенографии пациентам с травмой голеностопного сустава [134, 135].

2. Решения о безопасности исключения определённого целевого состояния. Пример: правило клинического решения в сочетании с анализом на D-димер для исключения тромбоза глубоких вен или тромбоэмболии лёгочной артерии [136].

3. Информирование будущих родителей о вероятности трисомии 21 у их будущего ребенка. Пример: тройные тесты во время беременности [137].

Возможные варианты клинического использования многофакторных прогностических моделей:

1. Информирование здоровых людей о 10-летнем риске развития сердечно-сосудистых заболеваний. Эту информацию можно использовать для изменения нездорового образа жизни. Примеры: шкала риска Framingham [138], QRISK2 [139], SCORE [140].

2. Информирование пациентов с диагностированным определённым заболеванием или подвергающихся определённой хирургической процедуре о риске неблагоприятного исхода или развития осложнения, чтобы определить превентивные меры или терапевтические стратегии. Пример: показания к тромболитической терапии на основании данных о 30-суточной летальности после острого инфаркта миокарда [141].

При разработке модели исследователи должны указать (в идеале на основе обзора литературы), были ли разработаны схожие модели (например, для такого же или подобного применения, участников или исходов) [47]. Исследования, посвящённые проверке моделей на независимых данных (*external validation studies*), дают ценную информацию об эффективности ранее разработанной модели у новых пациентов. Авторы должны ясно и однозначно указать, какую именно модель они проверяют со ссылкой на статью, а также указать или переформулировать (потенциальное) клиническое использование этой модели. Если существуют и другие конкурирующие предсказательные модели, авторы должны отметить, почему они оценивали только выбранную модель. Очевидно, что проверочное исследование, в котором сравнивают несколько конкурирующих моделей [48] на одних и тех же данных, предоставит дополнительную информацию [47, 85]. Необходимо также сообщить о любом запланированном изменении выборки, предикторов или исходов, по сравнению с исследованием, в котором была разработана модель (пункт 12), с обоснованием своего выбора.

Недавний систематический обзор проверочных исследований, основанных на внешних данных, показал, что авторы 7 из 45 (16%) работ не упомянули оригинальное исследование, в котором разрабатывалась оцениваемая предсказательная модель [122].

Пункт 3б. Укажите цели исследования, упомянув, идёт ли речь о разработке и (или) проверке модели (Р; П).

Примеры

«Цель исследования — разработать и проверить правило клинического прогнозирования для женщин с симптомами поражения молочной железы для внедрения в клиническое руководство более обоснованного подхода направления (к специалисту. — *Примеч. ред.*), включая срочное направление в соответствии с правилом двух недель» [142]. (Диагностика; Разработка; Проверка.)

«В этой статье мы сообщаем об оценке и внешней проверке новой параметрической прогностической модели, основанной на данных из Великобритании, для предсказания долгосрочной безрецидивной выживаемости пациентов с ранней стадией рака молочной железы. Эффективность модели сравнивали с показателями Nottingham Prognostic Index и Adjuvant Online. Представлены алгоритм подсчёта баллов и загружаемая программа для облегчения его использования» [143]. (Прогнозирование; Разработка; Проверка.)

«Широко признано, что никакая предсказательная модель не должна применяться на практике до формальной проверки её предсказательной точности (*predictive accuracy;* способность модели правильно классифицировать исход. — *Примеч. ред.*) у новых пациентов. Однако ранее ни в одном исследовании не проводилась формальная количественная (внешняя) проверка этих предсказательных моделей на независимой популяции пациентов. Поэтому сначала мы провели систематический обзор, чтобы идентифицировать все существующие модели, разработанные для предсказания продолжительности пребывания в отделении интенсивной терапии после кардиохирургического вмешательства. Впоследствии мы проверили эффективность обнаруженных моделей на большой независимой когорте кардиохирургических пациентов» [46]. (Прогнозирование; Проверка.)

Пояснение

Цели исследования — это конкретные задачи или исследовательские вопросы, которые будут рассматриваться в ходе исследования. Ясно и однозначно формулируя цели, часто в конце введения, авторы предоставляют

читателю необходимую исходную информацию, которая поможет критически оценить исследования. Для исследований предсказательных моделей цели должны указывать назначение предсказания (диагностическое или прогностическое), предсказываемые исходы или типы исходов, условия наблюдения и планируемую популяцию, для которой будет использоваться модель, а также тип предикторов, которые будут учитываться. Кроме того, авторы должны указать, касается ли отчёт разработки новой модели и (или) внешней проверки существующей модели.

Методы

Источник данных

Пункт 4a. Опишите дизайн исследования или источник данных (например, данные рандомизированного или когортного исследования, регистра), отдельно для наборов данных, использованных для разработки и проверки модели, если применимо (Р; П).

Примеры

«Популяционная выборка, использованная для данного отчёта, включала 2489 мужчин и 2856 женщин в возрасте от 30 до 74 лет на момент их обследования в Framingham Heart Study в период с 1971 по 1974 г. Участники приняли участие или в 11-м обследовании исходной когорты этого исследования, или в первичном обследовании дочернего проекта Framingham Offspring Study. В каждом случае использовали похожие протоколы исследования; лиц с явной ишемической болезнью сердца, обнаруженной в ходе первоначального обследования, исключали из состава участников» [144]. (Прогнозирование; Разработка.)

«Данные многоцентрового международного клинического испытания (*trial*) ADVANCE (Action in Diabetes and Vascular disease: preterax and diamicron MR controlled evaluation) позволяют получить новые уравнения для предсказания риска развития сердечно-сосудистых событий у пациентов с диабетом. <...> ADVANCE — факторное (*factorial*) рандомизированное контролируемое исследование контроля артериального давления (периндоприл/индапамид в сравнении с плацебо) и гликемии (интенсивное вмешательство на основе гликлазида MB в сравнении со стандартным лечением) на возникновение микро- и макрососудистых событий у 11 140 пациентов высокого риска с сахарным диабетом 2-го типа. <...> DIABHYCAR (The non-insulin-dependent DIABetes, HYpertension, microalbuminuria or proteinuria, CARdiovascular events, and Ramipril) — клиническое испытание рамиприла у лиц с сахарным диабетом 2-го типа, проведённое в 16 странах в период с 1995 по 2001 г. Из 4912 рандомизированных участников, 3711... подходили для проверки модели. Определения сердечно-сосудистых заболеваний в DIABHYCAR были похожи на таковые в ADVANCE. <...> В качестве предикторов учитывали возраст при постановке диагноза диабета, продолжительность заболевания, пол, ... режим лечения в соответствии с порядком рандомизации (снижение артериального давления и концентрации глюкозы в крови)» [145]. (Прогнозирование; Разработка; Проверка.)

«Провели многоцентровое проспективное проверочное исследование (*validation study*) с участием взрослых и наблюдательное исследование с участием детей, поступивших с острой травмой локтевого сустава в 5 отделений неотложной помощи на юго-западе Англии (Великобритания). Поскольку диагностическая точность теста (*diagnostic accuracy;* здесь и далее — правильность классификации целевого состояния. — *Примеч. ред.*) у детей не оценивалась, интервенционное исследование в этой группе не проводили» [146]. (Диагностика; Проверка.)

«Провели масштабную международную проверку индекса ADO, чтобы определить точность предсказания летального исхода у отдельных лиц с хронической обструктивной болезнью лёгких (ХОБЛ) в различных условиях и обновить индекс, если потребуется. Исследователи из 10 популяционных когортных исследований ХОБЛ, проводившихся в Европе и Америке, согласились сотрудничать в составе Международной рабочей группы» [147]. (Прогнозирование; Проверка; Обновление.)

Пояснение

Для разработки или проверки предсказательной модели можно использовать различные источники данных (*data sources*) или схемы исследований (*study designs*) (здесь эти термины используются как синонимы). Подробное описание дизайна, того, как были набраны участники исследования и собраны данные, предоставляет необходимые сведения о качестве данных, о том, был ли проведён надлежащий статистический анализ, и о возможности внешней обобщаемости (*generalizability*, экстраполяции результатов) предсказательной модели. Уязвимость к действию систематических ошибок варьирует в зависимости от дизайна исследования.

Диагностические исследования изучают одномоментную связь между диагностическими предикторами (характеристиками пациентов и результатами исследуемого теста) и наличием или отсутствием исхода (целевое состояние, представляющее интерес) (**вставка А**). Очевидный дизайн в этом случае — одномоментное исследование (*cross-sectional study*). В таких исследованиях набирают группу пациентов с определёнными характеристиками, у которых «подозревают наличие целевого состояния, представляющего интерес» [148–151]. Часто регистрация исхода (результатов референсного теста) происходит через некоторый промежуток времени после измерения предикторов. В идеале этот интервал должен быть как можно короче без начала какого-либо лечения в этот период. Из-за этого короткого периода времени и из-за того, что выбирается группа пациентов со схожими характеристиками (когорта), ведутся споры о том, следует ли обозначать эти исследования как чистые («*pure*») одномоментные исследования, или предпочтительно использовать термины «диагностические» (одномоментные)

когортные исследования («*diagnostic» cohort studies*) или «отложенные» («*delayed*») одномоментные исследования [152–154]. Если интервал между измерением предикторов и наступлением исхода продолжительный, и, безусловно, в случае начала лечения в этот период возникает опасность искажений в результате того, что статус (характеристики) заболевания у некоторых пациентов может измениться, тем самым изменится и одномоментная связь между предикторами и исходом.

В некоторых диагностических исследованиях сначала выполняют референсный тест (*reference standard*), а в исследование включают все случаи (пациентов с целевым состоянием), а также случайную выборку не-случаев (контролей). В таких исследованиях необходима корректировка частоты событий в общей выборке для получения несмещённых абсолютных вероятностей наличия диагностического исхода [152, 155–157]. К таким альтернативным схемам формирования выборки прибегают, когда распространённость исходов (целевых состояний) низкая, а затраты на измерение исследуемых предикторов или проведение исследуемого теста (*index tests*) высокие. Основной вопрос здесь: является ли такая выборка случаев и контролей репрезентативной в отношении целевой популяции, т.е. для пациентов, у которых только подозревают наличие целевого состояния. Явное нарушение происходит в исследованиях с неодинаковым отбором типичных, запущенных случаев и условно здоровых контролей [152, 155–157]. Такой отбор участников может привести к переоценке клинической значимости (*clinical relevance*) исследования [158], а многие показатели предсказательной эффективности (*predictive performance*) часто оказываются неверными [157].

Типичный дизайн прогностических исследований — продольное когортное исследование (*longitudinal cohort study*), которое может быть проспективным (*prospective*) или ретроспективным (*retrospective*) (**вставка А**) [1–3, 58, 103]. Участников включают в когорту на основании определённых критериев, таких как наличие конкретного заболевания, выполнение определённых хирургических процедур или беременность. Часто этот момент времени отмечают как T=0, исходный момент *(baseline)* или стартовая точка (*start point*) [9]. Затем за участниками ведётся наблюдение на протяжении некоторого периода времени, чтобы определить, развиваются ли у них интересующие исследователей события (исходы).

Предпочтительный дизайн — проспективное продольное когортное исследование. В этом случае полностью контролируется определение всех потенциальных предикторов и исходов (**рис. 2**), применяется наилучший метод для измерения каждого из них, при этом сводя к минимуму количество отсутствующих данных и выбывших из-под наблюдения участников.

Во многих исследованиях модель разрабатывают или проверяют с использованием набора данных, изначально собранных для других целей. Несмотря на то что исходное исследование изначально могло быть проспективным продольным когортным, конкретные предикторы могли не оцениваться его авторами, или некоторые предикторы могли быть оценены недостаточно хорошо. Пункт 13б указывает на необходимость представления детальной информации о количестве отсутствующих значений потенциальных предикторов, а пункт 13а — лиц, выбывших из-под наблюдения.

Рандомизированные испытания (*randomized trials*) — это особая подгруппа проспективных продольных когортных исследований, которые также могут быть использованы для разработки и проверки прогностических моделей. Однако здесь авторы должны указать, каким образом был учтён эффект вмешательства (пункт 5в). При использовании данных рандомизированных испытаний обобщаемость (*generalizability*) разработанной или проверенной модели может вызывать вопросы из-за многочисленных критериев невключения [1]. В одном исследовании было установлено, что прогностический эффект новых биомаркеров сердечно-сосудистых событий (добавленных к традиционной фрамингемской шкале риска, *Framingham risk score*) был сильнее в наборах данных, полученных в наблюдательных исследованиях (*observational studies*), чем в данных рандомизированных испытаний [159].

В условиях, когда международное сотрудничество и обмен данными становятся всё более распространённым явлением, для разработки и проверки предсказательных моделей всё чаще используются данные отдельных участников (*individual participant data*), полученные из многочисленных исследований [89, 147, 160]. Аналогичным образом используются существующие наборы данных большого объёма [так называемые большие данные (*big data*) из национальных или международных исследований или регистров] [139, 161, 162]. Данные из таких источников следует рассматривать как кластерные, поскольку участники происходят из разных кластеров (разных когорт, исследований, лечебных учреждений, условий наблюдения, регионов или стран), что требует взвешенного подхода при разработке предсказательных моделей. Недавно были предложены метааналитические подходы, учитывающие кластерную организацию таких данных [163–166]. Они учитывают разные наборы случаев с различным преваленсом (*prevalence*) (для диагностики) или инцидентностью исхода (*incidence*) (для прогнозирования) в когортах, наборах данных, исследованиях, больницах, условиях наблюдения, регионах или странах и, таким образом, учитывают различные исходные вероятности или риски (например, с помощью случайных свободных коэффициентов, *random intercepts*). Они также учитывают разные наборы случаев, отражающие разные ассоциации предикторов и исходов за счёт придания предикторам случайных весов (коэффициентов регрессии) [163–167]. Использование данных отдельных участников или источников больших данных расширяет возможности разработки и прямой оценки (внешней

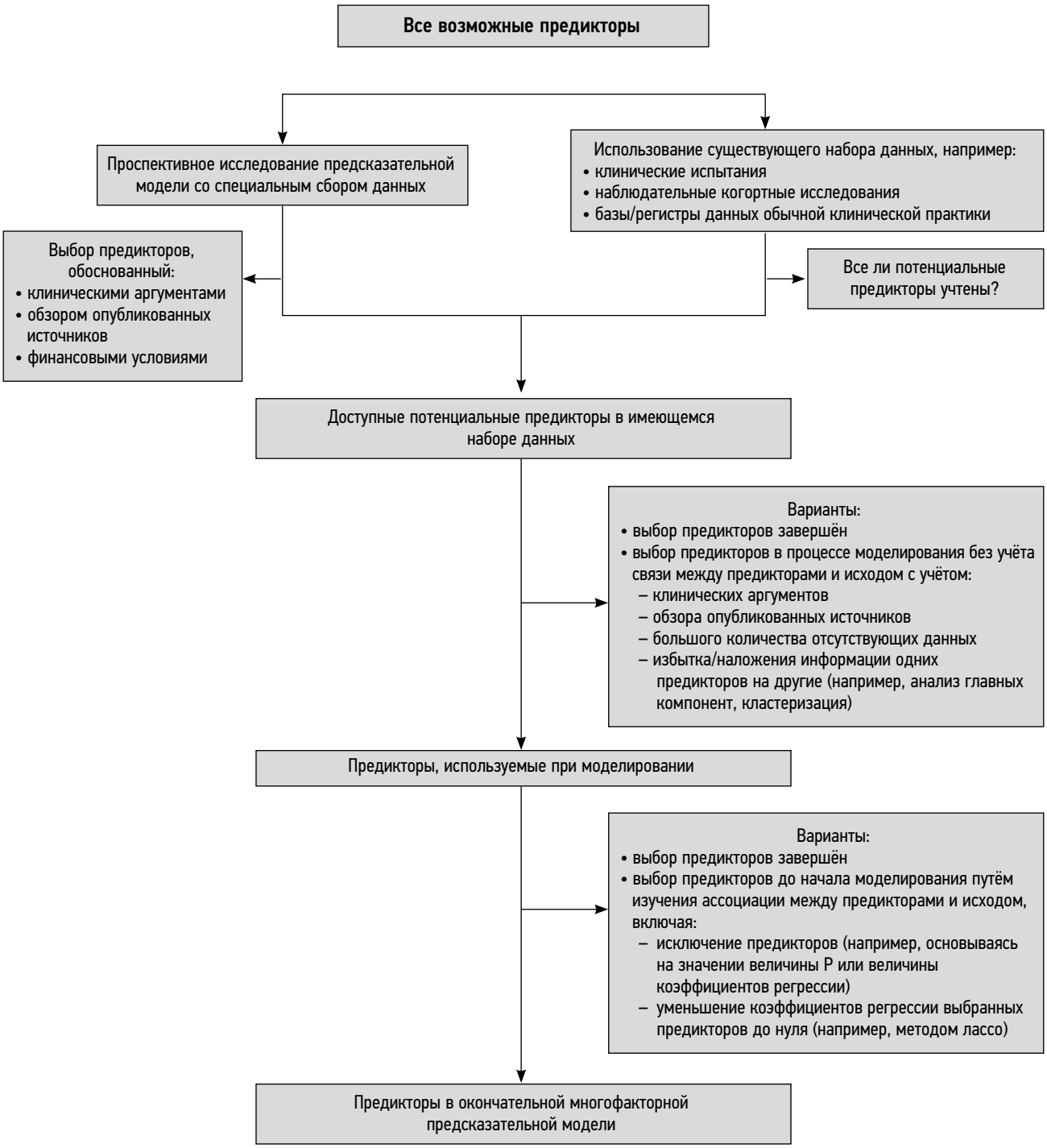


Рис. 2. Выбор предикторов в исследовании, посвящённом разработке многофакторной предсказательной модели.

проверки) предсказательных моделей в разных лечебных учреждениях, странах или условиях применения (**рис. 1**, исследование типа 2б), опять же с учётом потенциальных различий в весах свободных коэффициентов (*intercept*) и предикторов [164, 166]. Также недавно были предложены дополнения к часто используемым показателям эффективности модели для учёта кластеризованных данных [167–171].

Из соображений эффективности или стоимости исследователи из всей когорты могут отбирать отдельную группу пациентов. Примерами являются исследование типа «случай — когорта» (*case-cohort*) или вложенное (гнездовое) исследование (*nested case-control*) [172]. В отчётах о таких исследованиях необходимо описывать способы формирования выборки, что позволит корректно оценить абсолютную вероятность наличия или развития целевого исхода [1, 103, 173–175]. Избирательный подход к выбору или исключению участников может поставить под сомнение репрезентативность выборки для общей популяции, в которой будет применяться модель, и негативно сказаться на воспроизводимости результатов предсказательной модели.

Описание дизайна исследования или источника данных также предоставляет ценные сведения об условиях и первоначальной цели сбора данных. Информация об условиях проведения исследования (*setting*) и критериях отбора (*eligibility criteria*) (пункт 5б) позволит читателям судить о возможности применения модели в интересующих их клинических условиях.

В систематических обзорах исследований предсказательных моделей было отмечено, что их авторы зачастую неясно указывали, является ли выборка репрезентативной для целевой популяции, в том числе были ли в неё включены все последовательно отобранные участники [34, 59, 84, 93, 176].

Пункт 6б. Укажите основные даты исследования, включая начало и завершение набора участников и, если применимо, завершение периода последующего наблюдения (Р; П).

Примеры

«В проспективное исследование с временно́й проверкой (*prospective temporal validation study*) включили всех пациентов, которых лечили в период с марта 2007 по июнь 2007 г. в 19 испытаниях I фазы в отделении разработки лекарственных препаратов Королевской больницы Марсдена (Саттон, Великобритания). <...> Все пациенты находились под проспективным наблюдением до 31 мая 2008 г.» [177]. (Прогнозирование, Проверка.)

«В исследование включали всех пациентов, последовательно обратившихся с болью в передней части грудной клетки (в качестве основной или дополнительной жалобы) в течение периода от 3 до 9 недель (средняя продолжительность — 5 недель) с марта по май 2001 г. <...> В период с октября 2005 по июль 2006 г. 74 врача общей практики федеральной земли Гессен (Германия) последовательно включили в исследование всех пациентов с болью в передней части грудной клетки (в возрасте 35 лет и старше, *n*=1249). Период включения в исследование длился 12 недель для каждой общей практики» [178]. (Диагностика; Разработка; Проверка.)

«В когорту для разработки модели (*derivation cohort*) включили 397 пациентов в возрасте 18 лет и старше обоих полов, последовательно госпитализированных в одно из четырёх терапевтических отделений госпиталя Donostia в период с 1 мая по 30 июня 2008 г. Критерии

Таблица 2. Пример представления основных характеристик исследования (Диагностика; Разработка; Проверка) [181]

Характеристики	Популяция из Швейцарии (*n*=201)*	Популяция из США (*n*=258)*
Период сбора данных	Декабрь 1999 г. — февраль 2000 г.	Январь — март 2002 г.
Дизайн исследования	Проспективное когортное	Проспективное когортное
Условия проведения	Университетская клиника первичной медицинской помощи, обслуживающая городское население численностью 150 000 человек в г. Лозанна (Швейцария)	Отделение неотложной помощи или амбулаторные пациенты, нуждающиеся в интенсивной терапии Университетского госпиталя по оказанию специализированной медицинской помощи в г. Сан-Франциско (Калифорния)
Критерии включения	Взрослые амбулаторные пациенты с гриппоподобным заболеванием, диагностированным лечащим врачом	Все взрослые пациенты с симптомами острой инфекции дыхательных путей (кашель, синусит, боль, заложенность носа / ринорея, боль в горле или лихорадка), возникшими в предшествовавшие 3 недели
Исход	Наличие гриппа A или B	Наличие гриппа A или B
Референсный тест	Посев	ПЦР
Наличие вируса гриппа	104 (52,8)	53 (20,5)
Мужчины	101 (50)	103† (40)
Средний возраст (диапазон), годы	34,3 (17–86)	38,8 (18–90)
Лихорадка	116 (58)	54 (21)
Кашель	186 (93)	235 (91)
Боль в горле	151 (75)	181 (70)
Миалгия	181 (90)	154 (60)
Ринит	163 (81)	185 (72)
Головная боль	169 (84)	190 (74)
Озноб/потливость	166 (83)	158 (61)
Слабость	184 (92)	197 (76)
Появление симптомов в течение <48 ч	106 (33)	45 (17)

Примечание. ПЦР — полимеразная цепная реакция. * Значения указаны в виде *n* (%), если не указано иное. † Рассчитано для 256 пациентов.

невключения не использовали. В следующем году, в период с 1 мая по 30 июня 2009 г., аналогичным образом мы набрали когорту для проверки модели (*validation cohort*): 302 пациента в возрасте 18 лет и старше обоих полов, последовательно госпитализированных в одно из четырёх терапевтических отделений госпиталя» [179]. (Прогнозирование; Разработка.)

Пояснение

Информация о времени начала и завершения периодов исследования, в течение которых производили набор участников, помещает исследование в исторический контекст. Это даёт читателям необходимую информацию о доступных в эти периоды методах диагностики и лечения, а также о возможности выявлять определённые предикторы с применением современных (для периода исследования) медицинских технологий. Периоды исследования в сочетании с общим числом участников могут указывать, насколько избирательным было включение пациентов в исследование. Авторам отчётов следует указывать количество включённых в исследование участников за определённый период времени (например, за год) (см. пункт 13a).

Как уже обсуждалось в пункте 4a, временной интервал между определением предиктора и исхода в большинстве диагностических исследований является небольшим. Однако при отсутствии надёжного референсного теста пациенты могут находиться какое-то время под наблюдением, что позволит точнее оценить наличие целевого состояния (*target condition*) на момент определения предикторов. В таких случаях авторы должны сообщить, какой максимальный или минимальный интервал между измерением предикторов и финальной оценкой наличия или отсутствия целевого состояния допускался.

В прогностических исследованиях (*prognostic modeling studies*) продолжительность наблюдения (*follow-up*) имеет решающее значение для интерпретации эффективности модели (см. пункты 6a и 13a). Продолжительность периода наблюдения после включения в исследование может быть одинаковой для всех участников. В таком случае необходимо указать длительность исследования. Часто период наблюдения для всех включённых пациентов завершается в определённое время (о котором должно быть сообщено). Далее должны быть приложены усилия для установления статуса участника на дату закрытия исследования. События, которые происходят позже этой даты, игнорируются.

Систематические обзоры исследований предсказательных моделей показали, что авторы не всегда указывают основные периоды исследования [43, 122, 176, 180]. Например, из 61 исследования, посвящённого разработке или проверке моделей прогнозирования рака молочной железы, лишь 13 (12%) содержали сведения (даты) о начале и завершении периода набора пациентов и завершении последующего наблюдения [43].

Участники

Пункт 5a. Опишите условия и место проведения исследования (например, учреждения первичной или специализированной медицинской помощи, общая популяция), указав количество и местонахождение участвующих центров (Р; П).

Примеры

«На основании ранее разработанного алгоритма предсказания рисков (QRISK1) мы предложили новую версию алгоритма ... QRISK2. Провели проспективное когортное исследование в большой популяции амбулаторных пациентов в Великобритании с применением тех же методов, что и при первоначальном анализе. Данные извлекали из электронной базы QRESEARCH (версия 19) (www.qresearch.org). Это крупная электронная база верифицированных данных первичной медицинской помощи, которая насчитывает 11 млн записей о пациентах, зарегистрированных в 551 клинике общей практики» [139]. (Прогнозирование; Разработка; Проверка.)

См. также **табл. 2.**

Пояснение

Подробное описание того, где и когда были набраны участники исследования, особенно важно для того, чтобы другие могли судить об обобщаемости (*generalizability*) и полезности (*usefulness*) моделей, а также для проведения дальнейших исследований (например, проверка или применение модели на практике). Вопросы «Где?» и «Когда?» касаются не только географического положения и календарного времени, но и условий, в которых собирали данные об участниках (например, первичная, вторичная, третичная, неотложная медицинская помощь или общая популяция), а также кому оказывали помощь (взрослым или детям). Одного предположения, что предсказательные модели могут быть воспроизведены в других условиях или другой целевой популяции, недостаточно [19, 26, 28, 33].

В разных условиях может быть разная структура случаев, что влияет на обобщаемость и точность классификации (*accuracy*) предсказательных моделей (см. пункт 4a) [182–187]. Термин «структура случаев» (*case mix*) относится к распределению предикторов, других актуальных характеристик участников или условий, распространённости (в случае диагностики) или частоте исходов (при прогнозировании), что может привести к различным статистическим связям (ассоциациям) между предикторами и исходами, потенциально влияющими на предсказательную точность модели. Например, хорошо известно, что предсказательная эффективность моделей, разработанных для условий вторичной медицинской помощи, обычно ниже, чем когда они применяются в условиях первичной медицинской помощи [21, 183, 188]. Возможно, так происходит потому, что врачи учреждений первичной медицинской помощи или семейные врачи выборочно направляют пациентов к узким специалистам вторичного или третичного звена здравоохранения (*tertiary care*).

В результате популяция таких пациентов имеет более узкий диапазон характеристик, бóльшая доля — с заболеванием на поздних стадиях и часто с более высоким риском наступления интересующих исследователей исходов [102, 189, 190].

Ещё одна особенность условий проведения исследования — это переносимость (*transportability*) характеристик предсказательных моделей, разработанных для взрослой популяции, на педиатрическую практику [102]. Например, были разработаны различные предсказательные модели для оценки риска послеоперационной тошноты и рвоты у взрослых, которых планировали оперировать под общей анестезией. При проверке на детях предсказательная способность моделей существенно снизилась [191].

В целом характеристики модели будут более обобщаемыми, если структура случаев новой популяции находится в пределах структуры случаев исходной популяции (*development population*), с использованием данных которой разрабатывали исследуемую модель [186]. Однако, как указано в пункте 10д (см. также **вставку В** и **табл. 3**), ранее разработанную в одних условиях модель можно скорректировать или обновить относительно других условий, чтобы улучшить её воспроизводимость (*model transportability*).

Рекомендуем авторам представить таблицу с кратким изложением основных характеристик исследования для исходной и любой другой выборки, данные которой использованы для проверки модели [192]. Это необходимо для того, чтобы дать читателю представление о любых различиях в структуре случаев и потенциальных последствиях имеющихся различий (пункт 5б). Кроме того, авторам исследований, посвящённых исключительно проверке моделей, рекомендуем представить сводную таблицу с описанием не только проверочной выборки, но и выборки использованной для разработки модели.

Систематический обзор 48 исследований по созданию или проверке моделей для прогноза развития сердечной недостаточности, выявил, что отчёты 10 (21%) исследований не содержали данных о количестве медицинских центров, участвующих в исследовании [180].

Пункт 5б. Опишите критерии отбора участников (Р; П).

Примеры

«С 1987 по 2002 г. в отделениях дерматологии университетских клиник Мангейма и Бенджамина Франклина в Берлине обследованы 192 пациента с лимфомами кожи. У 86 человек диагностировали Т-клеточную лимфому кожи (ТКЛК) согласно классификации Европейской

Таблица 3. Обзор различных подходов к обновлению существующей предсказательной модели*

№	Метод обновления	Причина обновления
0	Без обновления (оригинальная предсказательная модель)	–
1	Корректировка свободных коэффициентов (исходный риск)	Отличия в частоте исходов (распространённость или инцидентность) между исходной выборкой, использованной для разработки модели (*development sample*), и выборкой для проверки модели (*validation sample*)
2	Метод 1 + корректировка всех коэффициентов регрессии предикторов по одному общему фактору (калибровочный коэффициент, *calibration slope*)	Чрезмерная (*overfitted*) или недостаточная (*underfitted*) подгонка коэффициентов регрессии или их комбинации в исходной модели
3	Метод 2 + дополнительная корректировка коэффициентов регрессии для предикторов с разным весом в проверочной выборке в сравнении с выборкой, использованной для разработки модели	Как и в пункте 2, + вес (коэффициента регрессии) одного или нескольких предикторов может отличаться в проверочной выборке
4	Метод 2 + выбор дополнительных предикторов (например, новых маркеров)	Как и в пункте 2, + один или более потенциальных предикторов не были включены в исходную модель, или может потребоваться добавление в исходную модель нового предиктора
5	Повторная оценка всех коэффициентов регрессии с использованием только данных проверочной выборки. При наличии данных выборки, использованной для разработки модели, оба набора данных могут быть объединены	Вес всех предикторов может отличаться в проверочной выборке, либо эта выборка значительно превосходит объёмом выборку, применённую для разработки модели
6	Метод 5 + выбор дополнительных предикторов (например, новых маркеров)	Как и в пункте 5, + один или более потенциальных предикторов не были включены в исходную модель, или может потребоваться добавление в исходную модель нового предиктора

Примечание. * Информация из источников [31, 290, 372, 373].

организации по исследованию и лечению рака. В соответствии с предложенной классификацией к основным типам ТКЛК относятся грибовидный микоз, синдром Сезари и другие редкие формы лимфом. <...> В исследование не включали пациентов с редкими типами ТКЛК, парапсориазом, псевдолимфомами и В-клеточными лимфомами кожи. <...> Стадирование ТКЛК осуществляли согласно классификации TNM (tumor-node-metastasis) Объединенной группы по грибовидным микозам. Синдром Сезари диагностировали у пациентов с признаками эритродермии и абсолютным количеством клеток Сезари в периферической крови >1000/мкл согласно критериям Международного общества по лимфомам кожи (ISCL)» [193]. (Прогнозирование; Разработка.)

«Критерии включения (*inclusion criteria*): возраст 12 лет и старше; травма, полученная в предыдущие 7 суток. Авторы выбрали возраст 12 лет как минимальное пороговое значение, поскольку отделение неотложной помощи принимает в основном пациентов 12 лет и старше, в то время как пациенты младшего возраста наблюдались в соседней детской больнице, находящейся примерно в полумиле от нашей больницы. В этом проведенное нами исследование отличалось от оригинальной работы д-ра Stiell, в которое не включали пациентов моложе 18 лет. Критерии невключения (*exclusion criteria*): беременность, спутанность сознания на момент осмотра, пациенты, направленные на рентгенографическое обследование, повторная госпитализация, множественные травмы, изолированные повреждения кожи (ожоги, ссадины, рваные и колотые раны)» [194]. (Диагностика, Проверка.)

Пояснение

Описание критериев отбора (*eligibility criteria*) важно для понимания потенциальной применимости и, значит, обобщаемости (*generalizability*) предсказательной модели. Авторы должны определить, кто мог стать или не стать участником исследования. Это необходимо для того, чтобы читатель получил представление, в отношении кого могут быть применены результаты и предсказания исследования.

Для проверочных исследований полезно сообщить, были ли критерии отбора для исследуемых аналогичны или отличались от тех, которые использовались в исходной модели. В приведённом выше примере [194] многофакторная диагностическая модель для выявления переломов голеностопного сустава, первоначально разработанная в Канаде, была проверена в Азии. Авторы описали детали отбора пациентов и сопоставили их с теми данными, которые использовались при разработке модели.

Если некоторые участники, соответствующие критериям отбора, не были включены из-за отсутствия данных (по предикторам или исходам), об этом следует сообщить. Исключение участников только по этой причине и ограничение анализа только теми, у кого есть такие данные, может привести к серьёзным систематическим ошибкам (*bias*) [195–201]. Ошибки могут возникать, если отсутствие данных носит не случайный, а намеренный характер (пункт 9).

Пункт 5в. Подробно опишите медицинское вмешательство, если применимо (P; П).

Пример

«Данные многоцентрового международного клинического испытания ADVANCE (Action in Diabetes and Vascular disease: preterax and diamicron-MR controlled evaluation) позволяют получить новые уравнения для предсказания риска развития сердечно-сосудистых событий у пациентов с диабетом. <...> ADVANCE — факторное (*factorial*) рандомизированное контролируемое исследование влияния контроля артериального давления (периндоприл/индапамид в сравнении с плацебо) и гликемии (интенсивное вмешательство на основе гликлазида MB в сравнении со стандартным лечением) на возникновение микро- и макрососудистых событий у 11 140 пациентов высокого риска с сахарным диабетом 2-го типа, отобранных в 215 медицинских центрах Азии, Австралазии, Европы и Канады. <...> В качестве предикторов учитывали возраст при постановке диагноза диабета, продолжительность заболевания, пол, систолическое, диастолическое и среднее артериальное давление, пульсовое давление, общий холестерин, липопротеины высокой плотности, липопротеины низкой плотности и триглицериды, индекс массы тела, окружность талии, отношение окружностей талии и бедер, лекарственные препараты для снижения артериального давления (лечение гипертензии), прием статина, курение в настоящее время, ретинопатию, фибрилляцию предсердий (в прошлом или в настоящем), логарифмически преобразованное отношение альбумина и креатинина в моче (ACR), креатинин в сыворотке крови (S_{cr}), гемоглобин A1c (Hb_{A1c}) и глюкозу в крови натощак, а также режим лечения в соответствии с порядком рандомизации (снижение артериального давления и концентрации глюкозы в крови)» [145] (Прогнозирование; Разработка; Проверка.)

Пояснение

Когорты для изучения прогноза определяются некоторым общим свойством здоровья [202]. Во многих прогностических исследованиях участники получают профилактические или лечебные вмешательства либо до, либо в начале периода наблюдения, что может повлиять на их прогноз. Эффективное лечение обычно благоприятно влияет на прогноз, что ведёт к снижению вероятности наступления изучаемого исхода [203].

Разработка чистой исходной (*pure baseline*) прогностической модели для предсказания будущих исходов у участников с определённым состоянием здоровья, которые не подвергались лечению, вряд ли возможна. Обычно участники получают некоторое лечение. В идеале либо все участники исследования получают одинаковое лечение, например хирургическое, либо методы лечения выбирают в результате рандомизации, например, если

прогностические модели основаны на данных рандомизированных испытаний (см. пункт 4а) [1, 204]. Некоторые прогностические модели специально разработаны и проверены для пациентов, получающих конкретное лечение [205], но даже здесь могут быть отличия в сопутствующих вмешательствах.

При использовании данных рандомизированных испытаний возможна раздельная разработка прогностических моделей для тех, кто получает различные варианты лечения, особенно при наличии эффективного вмешательства. Кроме того, методы лечения могут выступать отдельным предиктором для модели, разрабатываемой на основе данных всех пациентов (пункт 7а); также может быть изучено взаимодействие между лечением и другими предикторами (пункт 10б), чтобы сделать разные предсказания при разных стратегиях лечения [1, 4]. В таком случае основное внимание уделяется не профилактическим или терапевтическим эффектам вмешательства, а их независимому вкладу в предсказание исхода. Однако во многих случаях предсказательное значение вмешательств незначительно по сравнению с такими важными предикторами, как возраст, пол и стадия заболевания [1], поэтому такой фактор, как лечение, зачастую исключают на этапе моделирования или не учитывают в процессе выбора предикторов.

Для нерандомизированных исследований (*nonrandomized studies*) характерны не только вариации в получаемом лечении. Серьёзное беспокойство вызывает влияние на выбор лечения для отдельных лиц тех же предикторов, которые включены в статистическое моделирование [206]. Как и в случае с данными рандомизированных исследований, лечение также можно рассматривать как предиктор при моделировании, но влияние на предсказательную модель лечения, которое само является результатом влияния других предикторов, оценить нелегко. Предыдущие комментарии относятся к лечению до начала периода наблюдения. Для лечения, начатого позднее, требуются очень сложные модели, которые редко применяются в исследованиях предсказательных моделей [207].

Совершенно иная ситуация возникает, если лечение на текущий момент используется в качестве замещающей (*proxy*) переменной других предикторов, например, применение антигипертензивных или снижающих концентрацию холестерина препаратов как замещающих переменных гипертензии или гиперхолестеринемии, соответственно, в моделях сердечно-сосудистого риска [17, 208]. Влияние такого подхода на эффективность предсказательных моделей ещё недостаточно изучена.

Принимая во внимание вышеизложенные соображения, при разработке и проверке предсказательной модели важно знать, какие вмешательства в отношении участников исследования могли изменить вероятность наступления изучаемого исхода [203] (пункт 13б).

Вопросы лечения менее актуальны в большинстве исследований диагностических предсказательных моделей, поскольку эти исследования имеют одномоментный дизайн (*cross-sectional design*), в котором предикторы и исход фиксируются в одно время (**вставка А**). Иногда однако допускается некоторый интервал времени между измерением предиктора и исхода (например, если оценка исхода частично основана на данных последующего наблюдения) [209]. В таком случае информация о любом лечении, полученном в период между моментом предсказания и оценкой исхода, имеет важное значение и должна быть сообщена.

Недавний обзор 21 шкалы сердечно-сосудистого риска показал, что вмешательства, влияющие на исход, не учитывались, а сообщения о предшествующем лечении были неполными [203].

Исход

Пункт 6а. Определите предсказываемый моделью исход, включив описание способов и сроков его регистрации (Р; П).

Примеры

«Исходами были случаи смерти от любой причины, случаи смерти от ишемической болезни сердца и острые коронарные события. Для определения этих исходов за когортой участников наблюдали в течение определённого периода времени с помощью различных методов, включая ежегодные телефонные интервью, обследование в участвующих центрах каждые три года, наблюдение в государственных больницах, участвующих в исследовании ARIC (Atherosclerosis Risk in Communities), анализ свидетельств о смерти, результатов опроса врачей, отчётов коронеров/судебно-медицинских экспертов, интервью с лицами, владеющими конфиденциальной информацией. Период наблюдения стартовал с момента включения в исследование (1987–1989) и продолжался вплоть до 31 декабря 2000 г. Летальные коронарные события включали случаи смерти среди госпитализированных и негоспитализированных пациентов. К острым коронарным событиям относили установленный или вероятный инфаркт миокарда, смерть в результате ишемической болезни сердца или операции на сердце (коронарное шунтирование, коронарная ангиопластика) у госпитализированных пациентов или наличие изменений на электрокардиограмме при обследовании, проводимом в когорте каждые три года. Классификация событий подробно описана в <...>» [210]. (Прогнозирование; Разработка.)

«Инфекцию мочевыводящих путей определяли при обнаружении $\geqslant 10^8$ колониеобразующих единиц (КОЕ) одного типа микроорганизмов на литр мочи, полученной естественным путём, при $\geqslant 10^7$ КОЕ/л в образце мочи, полученном при катетеризации мочевого пузыря, или любой рост микроорганизмов в образцах надлобковой аспирации мочевого пузыря. Инфекцию мочевыводящих путей считали вероятной при $\geqslant 10^7$ КОЕ/л одного микроорганизма в моче, полученной естественным путём, при $\geqslant 10^6$ КОЕ/л одного микроорганизма в моче, полученной при катетеризации

мочевого пузыря, ⩾10^8 КОЕ/л микроорганизмов двух типов в моче, полученной естественным путём, или ⩾10^7 КОЕ/л микроорганизмов двух типов в моче, полученной при катетеризации мочевого пузыря» [211]. (Диагностика; Разработка; Проверка.)

«Для определения клинического исхода извлекали данные из медицинских карт пациентов и записей врачей. Пациентов, как правило, наблюдали в послеоперационном периоде не реже одного раза в 3–4 месяца в течение первого года, раз в полгода в течение второго и третьего года, затем ежегодно. Последующие обследования включали рентгенографию и компьютерную томографию для всех пациентов. Помимо физикального обследования и лабораторных анализов, по показаниям проводили внутривенную пиелографию, цистоскопию, цитологическое исследование мочи, уретральных смывов, остеосцинтиграфию. Локальный рецидив определяли как рецидив в хирургическом отделении, отдалённый рецидив — рецидив в других учреждениях после выписки. Клинические исходы отслеживали начиная от даты цистэктомии до даты первого зафиксированного рецидива, выявленного посредством компьютерной томографии, даты смерти или даты последнего обследования, когда у пациента ещё не было рецидива заболевания» [212]. (Прогнозирование; Разработка.)

«Выявление рака молочной железы: случаи рака молочной железы устанавливали путём опроса раз в два года в период с 1997 по 2005 г. О смерти членов семей узнавали из данных Почтовой службы США и Национального реестра смертности. Выявили 1084 случая рака молочной железы, из них 1007 случаев (93%) подтверждены медицинскими записями или данными регистра рака в 24 штатах, в которых проживали 96% опрошенных на момент включения в исследование» [213]. (Прогнозирование; Проверка.)

Пояснение

Исходы в диагностических моделях — это наличие или отсутствие конкретного целевого состояния (*target condition*) в момент времени T0 (**вставка А)**. Такие диагностические исходы определяют с помощью так называемого референсного теста (*reference standard*), т.е. самого доступного и общепринятого метода для установления наличия либо отсутствия целевого состояния [214]. Выбор такого метода должен быть обоснован. Референсный метод может иметь множество форм [отдельный тест, сочетание тестов или другой метод, включая консенсус экспертов или уполномоченной комиссии (*outcome committee*)]. Референсными тестами могут быть лабораторные, рентгенологические, артроскопические, ангиографические или патоморфологические исследования.

Если применимо, следует указать методы взятия проб крови или мочи, лабораторные и лучевые методы, технологии, а также определения, включая любые пороговые значения, которые использовали для оценки наличия (или тяжести) целевого состояния, и кроме того, правила, по которым объединяли результаты тестов (смешанный референсный тест) для установления диагностического исхода [215–217]. Если стандартные определения и пороговые значения не использовались, необходимо сообщить об этом с указанием причин. Если наличие/отсутствие исхода оценивали несколько исследователей (например, путём достижения консенсуса экспертной группой), необходимо описать метод установления окончательного диагноза (например, решение большинством голосов) [215, 216].

В исследованиях диагностических моделей необходимо указывать интервал времени между оценкой предикторов и исхода, поскольку изменения в состоянии пациента, которые могут произойти в этом интервале, могут быть причинами систематической ошибки (**вставка А**). Кроме того, должна быть однозначно описана последовательность оценки предикторов и исхода (см. пункт 6б и 7б о возможных систематических ошибках в связи с осведомлённостью исследователей, осуществляющих оценку).

В идеале диагностические исходы верифицируются у всех участников с использованием одного и то же референсного теста (*reference standard*). Но это не всегда возможно. Например, может считаться неэтичным применять инвазивный референсный тест при отсутствии положительного результата одного или более исследуемых тестов (*index tests*). В этом случае возможны два варианта: отложенная верификация (*delayed verification*), когда данные об исходах у участников, по которым не получены результаты референсных тестов, полностью отсутствуют (пункт 9), и дифференцированная верификация (*differential verification*), когда пациентов, которые не подвергаются предпочтительному референсному тестированию, оценивают посредством альтернативного референсного теста, отличающегося, как правило, более низкой точностью классификации [218, 219].

Например, в исследованиях, посвящённых диагностике рака, результаты патоморфологических (референсных) тестов, вероятно, будут получены только по тем участникам, у которых есть хотя бы один положительный результат исследуемого диагностического теста. Для остальных участников альтернативным референсным стандартом диагностики может быть период последующего наблюдения, достаточный для того, чтобы онкологические заболевания, имевшиеся на момент проведения исследуемого диагностического теста, стали очевидными (отложенная верификация), но не слишком продолжительный, иначе это может привести к выявлению новых случаев рака. Правила и процедуры, которые применяли при отложенной или дифференцированной верификации исходов, должны быть подробно описаны, что позволит оценить риск систематических ошибок, связанных с такими проверками (*partial/differential verification bias*) [156, 218–220]. Также следует сообщить о методах, которые применяли для корректировки таких систематических ошибок [219].

Для прогностических моделей часто анализируемыми исходами являются случаи смерти (от любой или конкретной причины), нефатальных осложнений или событий (например, инфаркт миокарда, рецидив рака, прогрессирование или начало заболевания), а также исходы, важные для отдельных пациентов (симптомы, функциональное состояние и качество жизни) [2]. Возможно также прогнозирование комбинации исходов. Например, целевое событие при изучении безрецидивной выживаемости (*disease-free survival*) в исследованиях злокачественных новообразований может включать локальные рецидивы, регионарные очаги, отдалённые метастазы и смерть (с регистрацией в качестве исхода того события, которое произойдет раньше) [221].

Все исходы должны быть однозначно определены. Если авторы используют стандартные определения (например, на основе Международной классификации болезней — МКБ) или их вариации, об этом следует сообщать и приводить ссылки на первоисточники. Технические детали, представленные в протоколе исследования или в предыдущих статьях, должны быть процитированы и в идеале быть доступными.

В прогностических исследованиях за участниками наблюдают в течение определённого периода времени и документируют время наступления целевого исхода после начала отслеживания таких событий (T0) (например, дата постановки диагноза или хирургического вмешательства) (см. **вставку А**). В некоторых исследованиях статус исхода оценивают у всех участников на протяжении фиксированного периода (например, общую выживаемость) и часто в заранее определённые моменты времени (например, 5- или 10-летний риск развития сердечно-сосудистых заболеваний), о чём также необходимо сообщать [222]. Аналогичным образом следует ясно указывать частоту оценки исходов во время периода наблюдения.

Следует сообщать об источниках данных, которые использовались для определения исходов или выбытия из-под наблюдения участников исследования (например, регистры случаев смерти, больничные записи, регистры злокачественных новообразований, клинические оценки, изображения или лабораторных анализы). Для таких исходов, как смерти вследствие определённых причин, процесс определения причины смерти необходимо ясно и однозначно описать (например, согласно решению экспертной комиссии, с обязательным кратким описанием состава и квалификации членов комиссии) [216].

Недавний обзор 47 исследований, в которых сообщалось о разработке моделей прогнозирования злокачественных новообразований, показал, что исходы были недостаточно ясно и однозначно определены в 40% исследований [54], в 30% случаев было неясно, являлась ли смерть исходом онкологического заболевания или регистрировались смерти от любой причины. Также наблюдались противоречия при описании событий, которые были включены в определение безрецидивной выживаемости.

Пункт 6б. Сообщите о любых действиях для маскирования (ослепления) при оценке предсказываемого исхода (P; П).

Примеры

«Все предполагаемые случаи серьёзной бактериальной инфекции рассматривались экспертной комиссией по окончательной диагностике, состоявшей из двух педиатров (с опытом работы в области инфекционных и респираторных заболеваний у детей), в случаях пневмонии дополнительно привлекали рентгенолога. Наличие или отсутствие бактериальной инфекции (исход) определяли путём маскирования экспертов в отношении клинической информации (исследуемые предикторы) и на основании консенсуса [211]. (Диагностика; Разработка; Проверка.)

«Биопсию печени выполняли иглой 18-го или большего калибра минимум с 5 портальными трактами с рутинным окрашиванием гематоксилин-эозином и трихромными красителями. Результаты биопсии интерпретировали в соответствии со схемой оценки, разработанной группой METAVIR, два эксперта в области патологии печени, ... которые не были проинформированы о клинических характеристиках и результатах исследования сыворотки пациентов. Оба эксперта оценили 30 биоптатов, согласованность заключений вычисляли, используя коэффициент каппа» [223]. (Диагностика; Разработка; Проверка.)

«Первичный исход (коронарная реваскуляризация в связи с острым инфарктом миокарда или смерть, наступившая вследствие сердечной или неизвестной причины, в течение 30 суток) подтверждён исследователями, которые не были осведомлены о прогностических переменных. Если установить диагноз не удавалось, кардиолог... изучал все клинические данные и устанавливал окончательный диагноз. Все положительные и 10% случайно отобранных отрицательных исходов подтверждены вторым исследователем, маскированным в отношении стандартизированных форм сбора данных. Разногласия разрешались консенсусом» [224]. (Прогнозирование; Разработка.)

Пояснение

В исследованиях предсказательных моделей исход в идеале должен оцениваться при сокрытии информации о предикторах. В противном случае эта информация может повлиять на оценку исхода, что ведёт к смещённым оценкам ассоциации между предикторами и исходом [148, 209, 225, 226]. Риск систематических ошибок будет меньшим при объективно (однозначно) измеряемых исходах (смерть по любой причине или, например, кесарево сечение). Однако он значительно возрастает при оценке исходов, требующих интерпретации (например, смерть вследствие конкретной причины).

Некоторые исходы трудно оценить вследствие их природы или по причине отсутствия общепринятых

референсных тестов. В таких случаях исследователи могут захотеть задействовать всю доступную информацию по каждому пациенту (включая данные о предикторах), чтобы определить наличие или отсутствие конкретного исхода. В диагностических исследованиях этот подход известен как *диагностический консенсус* (*consensus diagnosis*), а примерами в прогностических или интервенционных исследованиях являются решения экспертных комиссий (adjudication/end-point committees) (пункт 6a) [149]. Если явной целью является оценка дополнительной ценности (*incremental value*) конкретного предиктора или сравнение эффективности конкурирующих моделей (например, при проверке нескольких моделей), важность маскирования при оценке исходов возрастает, позволяя предотвратить переоценку дополнительной ценности предикторов или предвзятый выбор модели.

Исследователи должны тщательно обдумать и ясно указать, какой информацией располагали эксперты, оценивавшие исходы, и, если уместно, подробно описать, какие действия применялись для их маскирования (ослепления, *blinding*). Однако систематические обзоры сообщают о частом отсутствии в исследованиях информации о проведении маскирования при оценке исходов [34, 227].

Предикторы

Пункт 7a. Опишите все предикторы, использованные при разработке многофакторной предсказательной модели, указав, как и когда они были измерены (Р; П).

Примеры

«По каждому пациенту извлекали следующие данные: пол, аспартатаминотрансфераза (МЕ/л), аланинаминотрансфераза (МЕ/л), отношение аспартатаминотрансфераза/аланинаминотрансфераза, общий билирубин (мг/дл), альбумин (г/дл), насыщение трансферрина (%), средний объём эритроцитов (мкм^3), количество тромбоцитов (×10^3/мм^3) и протромбиновое время (сек). <…> Все лабораторные исследования проводились в течение 90 суток до биопсии печени. В случае неоднократных тестов учитывали результаты, наиболее близкие к моменту биопсии. Данные, полученные после биопсии, не учитывались» [228]. (Диагностика; Разработка.)

«Помимо возраста и пола, в модель предсказания смерти при остром инфаркте миокарда (ОИМ) включили ещё 43 потенциальных предиктора. <…> Эти переменные были взяты из списка факторов риска, которые использовались при разработке предыдущих бланков отчётности в проектах California Hospital Outcomes и Pennsylvania Health Care Cost Containment Council. Каждому сопутствующему заболеванию присваивали код по классификации МКБ-9 (Международная классификация болезней 9-го пересмотра) из 15 вариантов сопутствующих диагнозов в базе данных пациентов с инфарктом миокарда округа Онтарио (OMID). В базе данных OMID информация закодирована по классификации МКБ-9, а не МКБ-9 КМ, как это принято в США, поэтому коды, принятые в США удаляли. Некоторые факторы риска, учитываемые в вышеупомянутых проектах, не имеют аналога кода МКБ-9 (например, подтипы инфаркта, раса) и поэтому не были включены в наш анализ. Рассчитали частоту каждого из 43 сопутствующих заболеваний, исключая из дальнейшего анализа то заболевание, распространённость которого не превышала 1%. Сопутствующие заболевания, которые, по мнению авторов, не являлись клинически правдоподобными предикторами смерти при ОИМ, также были исключены» [185]. (Прогнозирование; Разработка; Проверка.)

«Каждый этап скрининга включал два посещения амбулаторного отделения с интервалом около трёх недель. Участники заполняли опросники, указывая демографические данные, сведения о наличии заболеваний сердечно-сосудистой системы и почек, курении, применении пероральных противодиабетических, антигипертензивных и липиднижающих препаратов. Информацию о применяемых лекарствах дополняли данными из розничных аптек, включая данные о классах антигипертензивных препаратов. <…> Во время первого и второго визита измеряли артериальное давление (АД) на правой руке каждую минуту в течение 10 и 8 мин соответственно автоматическим прибором Dinamap XL серии 9300 (Johnson & Johnson Medical Inc., Тампа, Флорида). Для систолического и диастолического АД учитывали среднее значение по двум визитам на этапе скрининга. Провели антропометрические измерения, взяли пробы крови натощак. Концентрацию общего холестерина и глюкозы в плазме крови измеряли стандартными методами. Креатинин сыворотки крови измеряли методом сухой химии (Eastman Kodak, Рочестер, Нью-Йорк) с коэффициентом внутрисерийной вариации 0,9% и межсерийной вариации 2,9%. рСКФ (расчётная скорость клубочковой фильтрации. — *Примеч. ред.*) оценивали по формуле исследования MDRD (Modification of Diet in Renal Disease) с учётом пола, возраста, расы и концентрации креатинина в сыворотке крови. Кроме того, участники собирали мочу в течение двух последовательных периодов длительностью 24 ч. Концентрацию альбумина в моче определяли методом нефелометрии (Dade Behring Diagnostic, Марбург, Германия), а величину экскреции альбумина с мочой — как среднее значение двух экскреций в пробах, взятых во время двух 24-часовых периодов. Объём потребляемых с пищей натрия и белка определяли по данным 24-часовой экскреции натрия и мочевины с мочой соответственно» [229]. (Прогнозирование; Разработка.)

Пояснение

Предикторы, как правило, определяют из числа демографических характеристик пациентов, данных анамнеза, физикального обследования, характеристик заболевания, результатов тестирований, предыдущего лечения [1]. Предикторы должны быть описаны максимально подробно с указанием единиц измерения любой количественной

переменной (*continuous predictors*) и всех категорий качественных переменных (*categorical predictors*), включая описание случаев объединения категорий. Это необходимо для того, чтобы читатели и другие исследователи при необходимости могли воспроизвести результаты исследования и, что более важно, проверить или применить предсказательную модель на практике. Если применимо, следует также указать методы формирования выборки исследования, методы лабораторной и визуальной диагностики, включая любые пороговые значения (*cut-offs*), по которым определяли наличие или (статистический) вес конкретного предиктора, или правила объединения предикторов (например, расчёт среднего артериального давления).

Авторы также должны объяснить, каким способом и когда определяли предикторы. Все предикторы должны быть определены до начала или в начале исследования, и их значения должны быть известны в тот момент, когда модель будет использована [1, 230, 231]. Образцы крови или тканей, собранные во время или до начала исследования, могут быть исследованы позже. Здесь главное — когда были получены образцы и использовались предикторы. Предикторы, которые определяли после начала исследования, целесообразно рассматривать как исходы, а не как предикторы, если только не используются методы обработки данных, учитывающие фактор времени [232]. Однако статистические методы обработки данных о предикторах, определяемых во время последующего наблюдения [233, 234], редко применяют в исследованиях предсказательных моделей. Методы определения предикторов (включая аналитические и лабораторные тесты) должны быть описаны достаточно полно и прозрачно с уровнем детализации, который позволит в последующем их воспроизвести, а также оценить обобщаемость предсказательной модели, включающей такие предикторы.

Во многих исследованиях, посвящённых разработке предсказательных моделей, собирают данные для большого количества предикторов, которые затем включают в статистический анализ (**рис. 2**). Однако чем больше предикторов учитывается, тем больше вероятность ошибочного включения слабых и малоинформативных предикторов в окончательную модель, что приводит к чрезмерной аппроксимации (*overfitting*) результатов или чересчур оптимистичным оценкам. Это особенно актуально для небольших наборов данных (см. пункт 8). Более того, модели с меньшим набором предикторов легче применять на практике, чем большие модели. Поэтому до или во время анализа данных часто бывает необходимо сокращать количество потенциальных предикторов [2, 235] (см. пункт 10б). Причины исключения любых предикторов из числа тех, которые учитывали на этапе моделирования, следует ясно и однозначно описать (**рис. 2**).

Недавно опубликованные систематические обзоры показали часто недостаточное описание всех доступных предикторов, общего количества проанализированных предикторов, способов и времени их отбора [34, 43, 45, 53, 54, 73, 74, 80, 81, 87, 182]. Согласно результатам обзора 29 предсказательных моделей в области репродуктивной медицины, 34% исследований не содержали надлежащего описания предикторов [80].

Пункт 7б. Сообщите о действиях для маскирования (ослепления) при оценке предикторов исхода или любых других предикторов (Р; П).

Примеры

«Количественный анализ изображений, полученных при магнитно-резонансном исследовании, выполнен одним исследователем, не осведомлённым о клинических данных пациентов и результатах эхокардиографии. (Цель заключалась в измерении дополнительной диагностической ценности (*incremental diagnostic value*) магнитно-резонансной томографии (помимо клинических данных), позволяющей подтвердить или исключить сердечную недостаточность.)» [236]. (Диагностика; Разработка; Дополнительная ценность.)

«Два сертифицированных врача скорой помощи, не информированные о (других) предикторах и исходах пациентов [нефатальные события и (или) случаи смерти, вызванные сердечно-сосудистыми заболеваниями; все случаи смерти в течение 30 суток после появления боли в грудной клетке] классифицировали все электрокардиограммы (один из исследуемых предикторов) в структурированном стандартизированном формате…» [224]. (Прогнозирование; Разработка.)

«Исследователи, не информированные о предикторах и исходах пациентов, проанализировали и классифицировали все электрокардиограммы в структурированном формате в соответствии с действующими стандартизированными правилами отчётности. Исходы оценивали два исследователя, которым не были предоставлены сведения из стандартизированных форм сбора данных. Исследователям были предоставлены результаты всех лабораторных анализов, рентгеновские снимки, данные ЭКГ с нагрузкой, катетеризации сердца, а также информация, полученная при телефонном опросе в течение 30 суток» [237]. (Диагностика; Проверка.)

Пояснение

Оценка предикторов может быть предвзятой, если исследователи осведомлены об исходах пациентов либо других предикторах [1, 225, 238–240]. Сокрытие информации (ослепление, маскирование) в отношении предикторов также важно, как и при оценке исходов (пункт 6б), особенно при оценке предикторов, основанной на субъективных суждениях (например, при интерпретации данных визуализирующих, электрофизиологических и морфологических исследований). Оценка таких предикторов, как пол, возраст или количественные данные лабораторных анализов, как правило, не зависит от интерпретации исследователя.

Сокрытие (ослепление) информации об исходе

Оценка предикторов всегда должна выполняться вслепую, без доступа к данным об исходах участников исследования. Данные об исходе будут ненамеренно включаться в оценку предикторов или искажать её, тем самым искусственно усиливая связь между предиктором и исходом [1, 225, 239]. Сокрытие информации (*blinding*) от исследователей, оценивающих предикторы, применяется в продольных исследованиях (*follow-up studies*), в которых исходы планируется измерять после предикторов, что присуще прогностическим исследованиям. Потенциальная систематическая ошибка вследствие оценки предикторов, выполненной при наличии информации об уже наступившем исходе, характерна для исследований «случай — контроль» и одномоментных исследований, где предикторы и исходы оцениваются в одном (близком) интервале времени [225]. Следовательно, такая систематическая ошибка с большей вероятностью будет наблюдаться в диагностических модельных исследованиях. Поэтому следует чётко указать, была ли доступна какая-либо информация об исходах при интерпретации результатов оценки предикторов (или результатов исследуемых тестов).

Сокрытие (ослепление) информации о других предикторах

Исследователи, оценивающие предикторы, могут иметь доступ к дополнительной информации (например, к данным анамнеза или медицинского осмотра). В отличие от сокрытия информации об исходе при оценке этих предикторов, сокрытие информации о других предикторах не является чем-то хорошим или плохим. Уместность сокрытия такой информации зависит от исследовательского вопроса и потенциальной значимости конкретных предикторов для клинической практики [209, 225, 226]. Интерпретация одних предикторов на основе ранее полученной информации о других предикторах может быть специально спланирована, если так происходит в повседневной практике. Например, предикторы из числа показателей, полученных при дополнительных исследованиях (методы визуализации, электрофизиологические исследования), как правило, интерпретируются с учётом данных анамнеза и физикального обследования.

Кроме того, если цель исследования заключается в оценке дополнительной ценности конкретного предиктора по отношению к предикторам, которые более-менее известны в клинической практике, ослепление эксперта, оценивающего конкретный предиктор, в отношении других предикторов крайне нежелательно. Однако если цель исследования — количественная оценка того, может ли конкретный предиктор или тест заменить другой предиктор или тест [например, может ли позитронная эмиссионная томография/компьютерная томография заменить традиционное сканирование лёгких для выявления раковых образований], ослепление исследователей, оценивающих эти два теста, в отношении результатов друг друга является необходимым во избежание предвзятости суждений [225, 239]. Без ослепления обе интерпретации, а значит, и полученные результаты будут более похожими.

Поэтому следует сообщить, какие оценки предикторов, если таковые имели место, не учитывали информацию о других предикторах в связи с целью исследования, а также где и как предикторы в модели будут использоваться на практике.

Многочисленные систематические обзоры показали, что слепая оценка (*blind assessment*) предикторов либо не выполняется, либо о ней не сообщается [3, 58, 67, 69, 95, 241]. Например, из 137 исследований, посвящённых разработке предсказательных моделей для целей педиатрии, лишь 47% содержали однозначные сведения о сокрытии данных о предикторах.

Размер выборки

Пункт 8. Объясните, как был определён размер выборки исследования (Р; П).

Примеры

«Мы определили размер выборки в соответствии с точностью оценки чувствительности разработанного правила принятия решения. Как и в предыдущих подобных исследованиях, мы предварительно определили 120 событий исхода, чтобы вывести правило с чувствительностью 100% с нижним 95% доверительным интервалом, равным 97,0%. Кроме того, чтобы добиться наибольшей полезности для практикующих врачей отделения неотложной помощи, мы планировали включить по крайней мере 120 событий исхода, произошедших вне отделения (в больнице или после выписки из отделения неотложной помощи). Обзор качественных данных госпиталя в Оттаве показал, что 10% пациентов, поступивших в отделение неотложной помощи с жалобами на боль в грудной клетке, соответствовали критериям исхода в течение 30 суток. По нашим оценкам, половина этих событий произойдет после госпитализации или выписки из отделения. Предварительно определённый размер выборки составил 2400 пациентов [224]». (Диагностика; Разработка.)

«Расчёт размера выборки выполняли, исходя из основной (первичной) цели (определить дополнительную по отношению к клиническим данным предсказательную ценность предоперационной КТ-ангиографии коронарных сосудов). Из двух наших целей для достижения именно этой требуется наибольшее количество пациентов для обеспечения стабильности предсказательной модели. <...> На основании данных пилотного исследования VISION и предыдущего неинвазивного исследования сердца, проведённого в аналогичной популяции, подвергавшейся хирургическим вмешательствам, мы ожидаем, что частота главных коронарных событий в операционный период составит 6%. В табл. 2 представлены различные размеры выборки, необходимые для проверки четырёх

переменных в многофакторном анализе, основанном на различной частоте и необходимом количестве событий для каждой переменной, включаемой в модель. Как видно из таблицы, при частоте событий 6% нам потребуется 1000 пациентов для получения достоверных оценок, при частоте 4% — 1500 пациентов. Мы ориентируемся на выборку из 1500 пациентов, хотя это количество может быть изменено в зависимости от частоты событий в выборке из 1000 человек [242]». (Прогнозирование; Разработка.)

«Использовали всю доступную в базе данных информацию, чтобы максимизировать мощность и обобщаемость результатов [243]». (Диагностика; Разработка.)

«Мы не производили формальных расчётов необходимого размера выборки, поскольку все когортные исследования (по теме исследования. — *Примеч. ред.*) ещё не завершены. К тому же не существует общепринятых подходов к определению размера выборки исследований, в которых выполняется разработка и проверка моделей предсказания рисков. Некоторые предложили при разработке модели иметь как минимум 10 событий исхода на одну переменную — кандидат на включение в модель, при проверке модели — 100 событий. Поскольку многие исследования по разработке и проверке предсказательных моделей являются небольшими, потенциальным решением будет крупномасштабное сотрудничество (как в нашем случае) для получения достоверных оценок на основе регрессионных моделей, которые, вероятно, могут быть распространены на другие популяции. Размер выборки и количество событий в нашем исследовании намного превосходят все значения, которые могут быть расчитаны при использовании существующих подходов к определению размера выборки, и потому мы ожидаем получить оценки высокой степени устойчивости [147]». (Прогнозирование; Проверка.)

«Мы рассчитали размер выборки исследования, необходимый для проверки предсказывающего клинического правила, в соответствии с потребностью включить не менее 100 пациентов с целевым исходом (наличие любой внутрибрюшной травмы), что подтверждается статистическими оценками, описанными ранее для внешней проверки (*external validation*) предсказывающего клинического правила. Исходя из нашей предыдущей работы, мы подсчитали, что частота (*prevalence rate*) внутрибрюшной травмы в исследуемой выборке составит 10%, и, таким образом, общий необходимый размер выборки должен составить 1000 пациентов» [244]. (Диагностика; Проверка.)

Пояснение

Хотя существует консенсус в отношении важности достаточного размера выборки для разработки предсказательной модели, неясно, какое именно количество участников следует считать достаточным. Если говорить о медицинских исследованиях в общем, то чем больше размер выборки, тем точнее будут полученные результаты. В отсутствие систематических ошибок крупномасштабные исследования также дают более надёжные результаты. Важно отметить, что при разработке и проверке предсказательных моделей эффективный размер выборки (*effective sample size*) обусловлен числом событий исхода. При бинарных (0/1) исходах или исходах, требующих учёта времени до события (*time-to-event outcome*; частный случай — анализ выживаемости. — *Примеч. ред.*), эффективный размер выборки определяется для исхода с меньшей частотой. Большой размер выборки может быть нецелесообразным, если исход зафиксирован лишь у нескольких пациентов.

Однако зачастую может быть доступен набор данных из крупных когорт с уже измеренными потенциальными предикторами и исходами. В таких случаях разумно использовать весь набор данных независимо от того, соответствует ли он расчётному размеру выборки. Авторы должны сообщить об этом, а не пытаться обосновать размер выборки на основе его произвольных апостериорных расчётов.

Разработка моделей (development study)

Как обсуждалось в пункте 10б, эффективность модели, вероятно, переоценивается, если при разработке и оценке предсказательной точности (*predictive accuracy;* здесь и далее — точность классификации предсказываемого события. — *Примеч. ред.*) использован один и тот же набор данных [23]. И эта проблема будет тем больше, чем меньше размер выборки [25, 32, 112]. Хотя оптимизм в отношении эффективности модели можно скорректировать путём внутренней проверки (*internal validation*) и применения методов коррекции (*shrinkage techniques*) (обсуждается в пункте 10б), предпочтительнее изначально иметь выборку большего размера. Эти опасения актуальны, даже если отбор предикторов не производится. Однако они серьёзнее, если такой отбор производят из большого числа доступных предикторов (**рис. 2**), особенно при отсутствии сильных. Для исследований с небольшой выборкой характерен высокий риск выбора ложных предикторов (*overfitting*; пункт 10б) и потери важных предикторов (*underfitting*) [25, 26, 32, 112].

На основе некоторых эмпирических исследований [245, 246] было предложено правило для расчёта размера выборки, которое получило достаточно широкое распространение. Правило заключается в том, чтобы иметь по меньшей мере 10 событий изучаемого исхода на каждую переменную модели (*events per variable, EPV*), а точнее, на каждый оцениваемый параметр. Некоторые, однако, заявили, что 10 событий — это слишком много [247], другие, что — слишком мало [25, 32, 248, 249]. Кроме того, возможно, что количество событий на одну переменную может быть не самым лучшим основанием для расчёта размера выборки [250]. В принципе, размер выборки должен быть таким, чтобы можно было с заданной точностью оценить определённые показатели

эффективности модели (*c*-индекс, R^2, оценка Бриера, чувствительность и специфичность и многие другие) [251–253].

На практике исследователи часто ограничиваются использованием доступного набора данных. Такие параметры, как количество событий на одну переменную, часто носят лишь описательный характер. Более того, контролировать этот показатель можно путём уменьшения количества анализируемых предикторов (**вставка В**). Предварительное определение размера выборки на основе статистических расчётов с использованием упомянутых выше подходов применимо только для проспективно планируемого исследования, посвящённого разработке предсказательной модели.

Авторы должны объяснить, как был определён размер выборки. Если статистически, то об этом следует сообщить со всеми подробностями. Часто размер выборки определяют, исходя из практических соображений, таких как время, доступность существующих данных или стоимость. В таких случаях полезно обсудить адекватность размера выборки по отношению к количеству исследуемых предикторов или основным показателям эффективности модели.

Проверка моделей (validation study)

Проверочные исследования имеют конкретную цель — количественную оценку эффективности существующей модели на основе других данных (**вставка В** и **рис. 1**). Определённых требований к размеру выборки для проверочных исследований нет, а эмпирических данных, которыми могли бы руководствоваться исследователи, недостаточно. Поэтому размер выборки часто обусловлен имеющимися данными, хотя в некоторых случаях может быть определён на основании статистических соображений.

Ограниченные эмпирические данные, которые могут помочь исследователям при выборе размера выборки для проверочных исследований, указывают на необходимость включения минимум 100 целевых (предсказываемых) событий и столько же случаев без этих событий [112, 254], предпочтительно более чем по 250 событий [2]. Однако эти предположения основаны на ограниченном количестве моделирований (*simulation studies*), применявших статистические схемы тестирования гипотез [например, достижение значений углового коэффициента (*calibration slope*) <1 или предварительно заданного снижения с-индекса], хотя предпочтительнее сконцентрировать внимание на точности (*precision*) получаемых оценок и правильности классификации (*accuracy*) при использовании новых данных.

Многочисленные систематические обзоры показали, что в исследованиях предсказательных моделей в случае как разработки, так и проверки моделей часто отсутствовало обоснование размера выборки или какое-либо описание проблемы выбора ложных предикторов [34, 54, 255].

Отсутствующие данные (missing data)

Пункт 9. Опишите, как обрабатывали отсутствующие (неполные) данные (например, анализ только полных наблюдений, подстановка значений), детально – применение любого метода подстановки значений (Р; П).

Примеры

«Мы предположили, что отсутствие данных носит случайный характер и зависит от клинических переменных и результатов коронарной ангиографии методом компьютерной томографии. Для подстановки применяли метод множественных импутаций (*multiple imputations)* с помощью связанных уравнений (*chained equations*). Отсутствующие значения предсказывали с помощью других предикторов с учётом данных коронарной ангиографии, выполненной методом компьютерной томографии, а также исхода. Мы создали 20 наборов данных, идентичных по известной информации, но отличающихся по значениям, использованным для подстановки пропущенных сведений, для отражения неопределённости, связанной с процедурой восстановления отсутствующих данных (*imputations*). Всего было добавлено 667 (2%) значений клинических данных. В нашем исследовании только небольшая часть пациентов прошла коронарную ангиографию методом катетризации. Анализ данных, ограниченный пациентами, перенёсшими такую ангиографию, мог показать искажённый результат вследствие систематической ошибки верификации/отбора (*verification bias*). Поэтому отсутствующие данные о результатах катетерной коронарной ангиографии восстанавливали, используя данные, полученные при коронарной ангиографии методом компьютерной томографии в качестве вспомогательной переменной в дополнение к другим предикторам. Результаты этих двух процедур хорошо коррелируют друг с другом, особенно при отрицательных результатах коронарной ангиографии методом компьютерной томографии. Сильная корреляция результатов указанных методов подтверждена на примере 1609 пациентов, прошедших обе процедуры (коэффициент корреляции Пирсона — 0,72). Результаты коронарной ангиографии методом компьютерной томографии не рассматривали в качестве предиктора, поскольку эти данные были использованы для заполнения отсутствующих данных. В случаях, если катетерная коронарная ангиография не была выполнена, в качестве переменной исхода использовали результаты коронарной ангиографии, выполненной методом компьютерной томографии (проведён анализ чувствительности, *sensitivity analysis*). Однако этот подход является более сложным, поскольку требует учёта других предикторов, а также неопределённости, связанной со значениями, используемыми для подстановки отсутствующих значений. Мы добавили 3615 (64%) значений исхода для катетерной коронарной ангиографии. Множественные импутации выполнили с использованием Stata/SE 11 (StataCorp)» [256]. (Диагностика, Разработка.)

«При отсутствии сведений об исходе данные о пациенте исключались из анализа. Для решения проблемы отсутствующих данных применяли метод множественных импутаций (*multiple imputation*) с помощью вызываемой посредством SAS программы IVEware (Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan, Анн-Арбор, Мичиган, США). Множественные импутации — проверенный и эффективный способ восстановления отсутствующих данных и минимизации систематической ошибки, которая часто может возникать в результате исключения таких пациентов. Кроме того, этот метод эффективен даже при большой доле отсутствующих данных. В модель множественной импутации включали значения четырёх исходов, возраст, пол, коды по классификации МКБ-9 Е, оценка по шкале комы Глазго в отделении реанимации, оценка по шкале комы Глазго во внебольничных условиях, оценка по шкале тяжести травмы, механизм травмы и вызовов бригады скорой помощи травматологического профиля. Создали 10 наборов данных с замещением отсутствующих значений, с последующим объединением значений площади под ROC-кривой по указанным наборам данных стандартным способом. Хотя общепринятого подхода к объединению ROC-кривых для наборов замещённых данных не существует, мы усреднили индивидуальные данные чувствительности и специфичности по 10 наборам, а затем нанесли эти точки на график для построения кривых в наших результатах» [257]. (Прогнозирование, Проверка.)

«Мы разделили данные на два набора: для разработки модели (обучающий набор) и её проверки (тестовый набор). Данные для разработки включали все случаи операций, выполненных за первые 5 лет, для проверки — все остальные сведения. Для обеспечения надёжности (*reliability*) данных мы исключили пациентов, по которым отсутствовала информация по основным предикторам: возраст, пол, последовательность операций, а также количество и локализация имплантированных сердечных клапанов. Кроме того, из набора данных, использованных для разработки модели, исключали сведения о пациентах, по которым отсутствовала информация о трёх и более оставшихся предикторах. Также в процесс моделирования не включали предикторы, значения которых наблюдали у менее 50% пациентов. Таким образом, были исключены такие показатели, как конечное диастолическое давление в левом желудочке, давление заклинивания в лёгочной артерии, градиент давления на аортальном клапане, наличие прогрессирующего эндокардита. Из тестового набора данных исключали пациентов, по которым отсутствовала информация по любому из предикторов, включённых в модель рисков. Чтобы выяснить, привело ли исключение пациентов из-за отсутствия данных к какой-либо систематической ошибке, мы сравнили ключевые предоперационные характеристики пациентов, исключённых из исследования, с теми, кто был в него включён. Отсутствующие значения всех остальных предикторов в обучающих данных были восстановлены с использованием метода множественной импутации. Было создано пять различных наборов восстановленных данных» [258]. (Прогнозирование, Разработка, Проверка.)

Пояснение

Почти все исследования предсказательных моделей сталкиваются с проблемой отсутствующих данных (*missing data*) по показателям исхода или предикторам. Однако далеко не во всех отчётах сообщается об этом, и ещё меньше исследователей пытаются решить эту проблему статистическими методами [34, 45, 53, 259]. В таких случаях разумно предположить, что участников с отсутствующими данными исключают из анализа, что приводит к так называемому анализу полных наблюдений (*complete-case analysis*). Включение только участников с полными данными не только неэффективно (это может значительно уменьшить размер выборки), но также может привести к смещению итоговых оценок, когда оставшиеся лица (строки статистических таблиц. — *Примеч. ред.*) со всеми необходимыми данными окажутся нерепрезентативными для всей исходной выборки исследования (т.е. будет сформирована особая подгруппа) (**вставка Г**). Для исследований, в которых разрабатывается или проверяется предсказательная модель, эта систематическая ошибка отбора (*selection bias*) приведёт к различным (смещённым) оценкам ассоциации предикторов с целевым исходом (при разработке модели) и предсказательной эффективности (при разработке и проверке модели), по сравнению с оценками, которые могли быть получены, если бы был проанализирован весь набор данных. Методы множественной импутации теперь встроены в наиболее часто используемые статистические пакеты (Stata, R, SAS), что позволяет восстанавливать любое отсутствующее наблюдение и выполнять последующий анализ множества наборов восстановленных данных. Мы отсылаем читателя к существующим рекомендациям по объединению целевых оценок (коэффициенты регрессии, показатели предсказательной эффективности) для исследований предсказательных моделей, выполненных с использованием методов множественной импутации (**вставка Г**).

Авторам исследований предсказательных моделей рекомендуется подробно описывать отсутствующие данные (пункт 13б) и действия в отношении таких данных (пункт 9). Если из анализа исключали лиц с отсутствующими данными, это должно быть ясно отражено в критериях отбора (пункт 5б) с обоснованием причины исключения.

Основные детали, которые следует включить в описание того, как обрабатывались отсутствующие данные согласно стандартным рекомендациям [56, 200, 259], представлены в **табл. 4**. Если исследование посвящено одновременно и разработке, и проверке предсказательной модели, авторы должны ясно указывать, как отсутствующие данные были обработаны для обоих наборов данных и описывать любые различия в этих действиях.

Систематические обзоры методологии проведения и отчётности исследований, в которых разрабатывались и оценивались предсказательные модели, неизменно демонстрируют низкое качество описания информации об отсутствующих данных и их обработке [34, 43, 45, 53, 56, 59, 60, 62, 64, 66, 70, 71,76, 78–84, 88, 93, 122, 176, 260, 261].

Методы статистического анализа

Пункт 10a. Опишите, как поступали с предикторами в процессе анализа данных (P).

Примеры

«После оценки нелинейности с использованием ограниченных кубических сплайнов (*restricted cubic splines*) было обнаружено, что линейная связь непрерывных предикторов возраста, глюкозы и Hb (гемоглобина) с исходом является хорошим приближением» [262]. (Прогнозирование.)

«С помощью дробных полиномов исследовали наличие нелинейной зависимости исхода (годы до наступления события) от непрерывных предикторов (возраст, ИМТ (индекс массы тела)» [258]. (Прогнозирование.)

«Нелинейная связь между этими предикторными переменными и риском развития рака лёгких оценивали с помощью ограниченных кубических сплайнов. Сплайны для возраста, количества пачек выкуриваемых сигарет в год, времени отказа от курения и продолжительности курения задавали, размещая точки интерполяции на основе процентильных распределений этих переменных только в группе курильщиков. Такими точками для возраста были 55, 60, 64 и 72 года; интенсивности курения — 3,25; 23,25 и 63 пачки в год; времени отказа от курения — 0, 15 и 35 лет; продолжительности курения — 8, 28 и 45 лет» [263]. (Прогнозирование.)

Вставка Г. Отсутствующие данные

Отсутствующие значения предикторов либо исходов характерны для всех типов медицинских исследований, в том числе посвящённых разработке и проверке диагностических и прогностических моделей. Большинство статистических пакетов для статистической обработки по умолчанию исключают лиц с отсутствующими значениями любых данных, включённых в анализ. Наиболее распространённым подходом к обработке отсутствующих данных является так называемый анализ доступных (*available case*) или полных наблюдений (*complete case*). Однако даже небольшое количество отсутствующих данных по каждой из переменных исследования может привести к исключению большого количества пациентов из многофакторного анализа. Простое исключение записей с отсутствующими данными не обязательно повлияет на достоверность полученных результатов, если удалённые записи представляют собой полностью случайное подмножество исходной выборки исследования [195–200]. Однако если лица с отсутствующими данными нерепрезентативны для исходной выборки исследования, анализ исключительно полных наблюдений будет подвержен систематической ошибке. Степень смещения будет зависеть от различных факторов, включая количество лиц с отсутствующими данными [10, 195–201, 492]. Показано, что использование отдельной категории, указывающей на отсутствующие данные, также приведёт к искажённым оценкам, и такой подход является клинически бессмысленным для исследований предсказательных моделей, и его следует избегать [195, 196].

Отсутствие данных описывается как полностью (однозначно) случайное (*missing completely at random*, MCAR), если вероятность отсутствия конкретного наблюдения не зависит от наблюдаемых переменных исследования, предикторов или исхода; как (просто) случайное (*missing at random*, MAR), если отсутствие данных связано с другими наблюдаемыми переменными; как неслучайное (*missing not at random*, MNAR), если вероятность этого зависит от ненаблюдаемых значений, включая, возможно, и само отсутствующее значение [493, 494]. И хотя можно проверить, является ли отсутствие данных полностью случайным или связанным с наблюдаемыми переменными, доказать, что такие данные относятся к категории MAR и тем более MNAR, невозможно.

Вместо того чтобы просто исключать всех лиц с отсутствующими данными или использовать метод введения индикаторных переменных (*missing indicator method*), более эффективной группой методов для решения проблемы отсутствующих данных, связанных с исследуемыми переменными, что позволяет предполагать наличие механизма MAR, являются так называемые методы импутаций (*imputation techniques*). Они включают подстановку (вместо отсутствующего) среднего арифметического значения или медианы, стратифицированную импутацию или подстановку в подгруппах или использование многофакторной модели. Последний подход может быть реализован путём восстановления отсутствующего значения единичным (*single imputation*) или несколькими значениями (*multiple imputation*) [493–495].

Метод множественных импутаций считается предпочтительным для восстановления отсутствующих данных, позволяя получить более корректные значения стандартных ошибок и величины *P*; при восстановлении отсутствующих данных единичным значением мы получаем слишком низкие значения этих статистических показателей, что завышает вероятность ложноположительных результатов [103, 195–200, 492, 496]. Множественная импутация предполагает создание нескольких копий набора данных, в которых отсутствующие значения заменяются восстановленными, полученными из их предсказанного распределения по наблюдаемым

данным [493, 497]. Стандартный текст о множественной импутации обычно предполагает, что 5 или 10 копий набора данных будет достаточно. Однако совсем недавно было высказано предположение, что количество импутаций должно быть намного бóльшим и соотносимым с долей отсутствующей информации в данных [495]. Наконец, к каждому набору восстановленных данных, которые могут быть объединены (по правилу Rubin [494]), может применяться стандартный статистический анализ. Это позволит получить общую оценку для каждого коэффициента регрессии или показателя эффективности модели (пункт 10г) [2, 498], принимая, таким образом, во внимание неопределённость в восстановленных значениях [196–201, 492, 495, 499, 500].

Пояснение

Многие предикторы представляют собой непрерывные переменные (*continuous variables*). Исследователи должны решить, как поступать с ними в процессе анализа. При разработке клинических предсказательных моделей часто прибегают к их преобразованию в категориальные предикторы (*categorical predictors*). Однако этот подход вызывает серьёзные опасения. Во **вставке Д** объясняется, почему в идеале непрерывные переменные должны анализироваться как есть, и необходимость проверки линейной или нелинейной связи с изучаемым исходом.

При отсутствии априорного клинического консенсуса авторам, желающим категоризировать или даже дихотомизировать непрерывные предикторы, рекомендуется использовать метод, не основанный на данных. Определённо следует избегать выбора так называемых оптимальных пороговых значений (*optimal cut points*) путём минимизации значения *P* [264, 265]. Такие основанные на данных подходы крайне несовершенны, что приводит к оптимистичным или ложным оценкам ассоциации между предикторами и исходами, что способствует переобучению модели и, следовательно, неточной (оптимистичной) оценке её эффективности.

Категориальные предикторы можно также обрабатывать разными способами до начала анализа данных. В частности, немногочисленные данные можно сгруппировать таким образом, чтобы все они относились к одной категории. Например, для редких гистологических типов (опухоли. — *Примеч. ред.*) можно создать отдельную категорию «Другие типы». Любые изменения категорий необходимо объяснить (см. также пункт 10б).

Вставка Д. Непрерывные предикторы*

Многие предикторы регистрируются как непрерывные измерения, но преобразуются для анализа в категориальную форму с использованием одной или нескольких пороговых значений (*cut points*) (пункт 10а) [501]. Это делается для того, чтобы упростить анализ, облегчить клиническим специалистам использование предикторов или предсказательной модели, поскольку ассоциация между предиктором и исходом часто неизвестна, или с целью графического описания (например, с помощью кривых Каплана–Мейера). Хотя для принятия клинических решений категоризация оценок вероятности предсказательных моделей необходима, важно понимать, что для статистического анализа не требуется проводить категоризацию непрерывных

Таблица 4. Основная информация об отсутствующих данных, которую необходимо включить в отчёты

В разделе «Методы»:

- Ясное и однозначное описание метода, использованного для учёта отсутствующих данных показателей предикторов и исхода [например, анализ полных наблюдений, восстановление отсутствующих данных единичным (*single imputation*) или несколькими значениями (*multiple imputation*)].
- Возможные причины отсутствия данных.
- Если проводился анализ восстановленных данных (путём единичных или множественных импутаций):
  - сообщите об использованном программном обеспечении (включая любые специальные процедуры восстановления, например ICE, MICE, PROC MI, Amelia, aregImpute);
  - перечислите переменные, включённые в процедуру восстановления данных, указав также, учитывались ли исходы при восстановлении значений предикторов, и наоборот;
  - объясните, каким образом при восстановлении данных обрабатывали непрерывные, бинарные и категориальные предикторы;
  - сообщите, включали ли какие-либо взаимодействия в модель восстановления данных;
  - укажите количество копий восстановленных данных, если применялся метод восстановления несколькими значениями.

В разделе «Результаты»:

- Количество лиц с отсутствием хотя бы одного значения (любого из исследуемых показателей. — *Примеч. ред.*), только одного значения, двух значений и т.д.
- Количество отсутствующих значений (по предикторам и показателю исхода).
- Сравнение характеристик лиц с отсутствующими значениями и тех, кто имел полный набор данных. Это позволит понять, является ли отсутствие данных по конкретным переменным исследования (предикторам или исходам) полностью случайным или связанным с наблюдаемыми характеристиками (**вставка Г**).

переменных, включаемых в модель. Как поясняется ниже, за очевидные на первый взгляд преимущества упрощённого анализа приходится платить высокую цену.

Категоризация

Категоризация (*categorization*) позволяет исследователям избегать слишком оптимистичных предположений о связи между предиктором и исходом. Однако происходит это за счёт потери информации. Очевидно, что потеря информации будет наибольшей в случае дихотомизации предиктора (две категории). Хорошо известно, что результаты (например, предсказывающая эффективность модели) могут отличаться, если для дихотомизации предиктора используются различные пороговые значения. Однако если пороговое значение выбирают на основе множественного анализа данных, в частности, при наименьшем значении *P,* тогда и значение *P* для этого предиктора будет крайне низким, а оценки эффективности модели — чересчур высокими (сверхоптимистичными) [264].

Даже при заранее заданном пороговом значении дихотомизация статистически неэффективна и крайне нежелательна [265, 502–505]. Более того, если пороговые значения необходимы для помощи в классификации людей по отдельным группам риска, это следует делать на основе предсказанных вероятностей или рисков [30, 265].

Разделение непрерывной переменной на три или более категории снижает потери информации, но это редко используется в клинических исследованиях. И даже в этом случае пороговые значения приводят к получению модели со ступенчатыми функциями, что некорректно описывает плавные связи (*smooth relationship*), существующие в действительности [266].

Сохранение непрерывности переменных

Линейная зависимость (*linear functional relationship*) — наиболее популярный подход для сохранения непрерывности предиктора. Часто это является приемлемым допущением, но оно может быть неверным, что приводит к построению некорректной модели, в которую может быть не включён важный предиктор или в которой предполагаемая связь между предикторами и исходами (*predictor-outcome relationship*) значительно отличается от неизвестной истинной связи. Проверку линейности связи можно выполнить, исследуя возможное улучшение характеристик модели при допущении некоторой формы нелинейности связей. В течение долгого времени для моделирования нелинейной зависимости использовали квадратичные или кубические полиномы, но более общее семейство дробных полиномов (*fractional polynomials*) позволяет получить обширный класс простых функций, которые часто обеспечивают лучшие характеристики модели [506]. Определение требований к дробным полиномам и выбор модели можно выполнять одновременно с простым и понятным представлением результатов [266, 297].

Сплайн-функции, в частности ограниченные кубические сплайны (*restricted cubic splines*) — это ещё один подход к исследованию зависимости непрерывных предикторов [112]. Ограниченные кубические сплайны рекомендуют вместо стандартных кубических сплайн-функций, поскольку последние часто демонстрируют неудовлетворительные результаты в хвостах распределения значений предиктора [112, 507]. Ограниченные кубические сплайны чрезвычайно гибки, но на сегодняшний день нет общепринятой процедуры одновременного выбора предикторов с определением формы связи. Более того, даже при построении моделей с использованием одномерных сплайнов представление результатов обычно сводится к построению графика зависимости исходов от предикторов, поскольку представление коэффициентов регрессии часто бывает слишком сложным.

* Текст этой вставки в значительной степени совпадает с текстом **вставки 4** в работе [108].

Авторы должны ясно и однозначно сообщить, как анализировали предикторы. В частности, следует обосновать (теоретически или клинически) разделение непрерывных переменных (*continuous predictors*) на категории, указать пороговые значения (*cut points*) и то, как они были выбраны. Для каждого предиктора, использованного в виде непрерывной переменной, авторы должны уточнить, были ли данные сохранены в исходных единицах или преобразованы (например, логарифмическое преобразование). Необходимо сообщить, какой тип связи между предикторами и исходами был смоделирован (линейная или нелинейная) с указанием метода, если связь рассматривалась как нелинейная (например, использованы дробные полиномы, *fractional polynomials*, или ограниченные кубические сплайны, *restricted cubic splines*). Если связь считали линейной, желательно сообщить, проверялось ли предположение о линейной связи с предсказываемым исходом.

Экстремальные значения также могут быть скорректированы к менее экстремальным значениям во избежание нежелательных последствий, связанных с эффектом рычага (*leverage effects*) [2, 266]. Авторы должны сообщить, как поступили с маловероятными наблюдениями (например, вносили изменения или не учитывали при анализе).

Хотя информацию об анализе предикторов, как правило, представляют в разделе «Методы», вместе с определениями категорий (пункт 7а) её можно включить и в таблицы раздела «Результаты» (пункты 13б и 13в).

Обзоры опубликованных исследований неизменно показывают, что категоризация непрерывных предикторов — довольно распространённое явление. При этом многие авторы выполняют дихотомизацию всех предикторов [34, 41, 43, 45, 53, 54, 62, 63, 267, 268]. Так, обзор 11 исследований, посвящённых изучению

аневризматического субарахноидального кровоизлияния, показал, что такой предиктор, как возраст, дихотомизировали для всех моделей [81]. Обзор моделей для предсказания исходов рака показал, что авторы 12 из 45 работ (30%) не пояснили принципы кодирования значений всех предикторов в окончательной модели [55]. Другие обзоры также показали отсутствие ясности в вопросах обработки непрерывных предикторов [54, 64].

Пункт 10б. Укажите тип модели, последовательность её построения (включая выбор предикторов) и методы внутренней проверки (Р).

Авторам следует описать все статистические методы, использованные при разработке предсказательной модели. Представленная информация должна быть достаточно подробной, чтобы осведомлённый читатель при доступе к исходным данным мог проверить полученные результаты (www.icmje.org). Более того, читатель должен понимать причины, по которым был выбран тот или иной подход к анализу данных.

При разработке предсказательной модели можно следовать множеству возможных стратегий анализа. Выбор делается на каждом этапе анализа [2, 112, 266, 269]. Некоторые решения по стратегии моделирования основываются на данных, а также на медицинском контексте. Например, может потребоваться разработка модели только с несколькими основными предикторами для повышения клинической применимости модели (пункты 3а, 19б и 20) в ущерб её прогностической эффективности.

Основная проблема многих исследований предсказательных моделей заключается в том, что может быть выполнено множество различных анализов, но сообщается только о наилучшей предсказательной модели (т.е. с наилучшей дискриминацией; *discrimination* — классификация наблюдений в соответствии с предсказываемым исходом. — *Примеч. ред.*) [1]. Такая избирательность, основанная на данных, может привести к выбору переобученной модели с чересчур оптимистичными характеристиками эффективности. Это станет очевидным, если оценить модель на основе нового набора данных, взятых из исходной популяции [270]. Именно поэтому авторам следует представлять исчерпывающие сведения о всём диапазоне выполненного анализа. При необходимости полное описание статистического анализа можно представить в приложении к статье с указанием использованного компьютерного кода (пункт 21). В идеале этот код должен быть представлен вместе с индивидуальными данными участников, что обеспечит полную воспроизводимость, хотя это может быть неосуществимо, если не будет согласован открытый доступ к данным [271].

В последующих разделах мы рассмотрим специальные аспекты анализа, выполняемого в ходе разработки моделей. Не все из них будут актуальны для некоторых исследований. Более подробное обсуждение методов статистического анализа, как бинарных, так и зависящих от времени исходов (*time-to-event outcomes*), можно найти в других источниках [2, 12, 112, 266, 272–277].

1. Тип модели

Примеры

«Мы использовали модель пропорциональных рисков Кокса в наборе исходных данных для оценки коэффициентов, ассоциированных с каждым потенциальным фактором риска, для впервые зарегистрированного диагноза сердечно-сосудистого заболевания отдельно у мужчин и женщин» [278]. (Прогнозирование.)

«Все клинические и лабораторные предикторы были включены в многофакторную модель логистической регрессии (исход — бактериальная пневмония)» [279]. (Диагностика.)

Пояснение

В исследованиях медицинского предсказания используют различные типы моделей [112]. Большинство моделей получают с использованием многофакторной регрессии (*multivariable regression*). Модель логистической регрессии чаще всего применяется для бинарных конечных точек (*binary endpoints*), таких как наличие или отсутствие заболевания в диагностических моделях или краткосрочные события в прогностических моделях (например, 30-дневная летальность). Полупараметрическая регрессионная модель пропорциональных рисков Кокса (*semi-parametric Cox proportional hazards regression model*) чаще всего применяется для прогнозирования наступления исходов во времени (как правило, долгосрочных, например, 10-летний риск развития сердечно-сосудистых заболеваний), хотя для таких целей подходят и полностью параметрические модели (*fully parametric models*) [280, 281].

Авторы должны чётко идентифицировать используемую регрессионную модель. Если для прогнозирования долгосрочных исходов используется метод логистической регрессии, такой выбор необходимо обосновать. Разработка (и проверка) моделей, предсказывающих долгосрочные исходы при помощи метода логистической регрессии, требует, чтобы все участники находились под наблюдением в течение всего периода отслеживания исходов.

Доступно множество вариантов регрессионных моделей для бинарных, полиномиальных, упорядоченных, непрерывных и других исходов [2]. Другие типы предсказательных моделей включают деревья регрессии и методы машинного обучения (*machine learning techniques*), такие как нейронные сети (*neural networks*) и метод опорных векторов (*support vector machines*) [275]. При таком альтернативном подходе авторам следует обосновать свой выбор.

2. Выбор предикторов до начала моделирования

Примеры

«Мы отбирали факторы риска, исходя из результатов предыдущих метаанализов и обзора, простоты их

использования в учреждениях первичной медико-санитарной помощи, модифицируемости или обратимости этих факторов посредством изменения привычек (например, курение) или терапевтического вмешательства. Однако наш выбор был ограничен факторами, которые уже были использованы в двух исходных когортах, составлявших базу данных EPISEM» [282]. (Прогнозирование.)

«К потенциальным предикторам (*candidate variables*) из каждого источника данных относили все демографические и связанные с заболеванием факторы, а также способы лечения, которые, как было показано, являются факторами риска смерти после предшествующего эпизода интенсивной терапии. Первоначальный отбор переменных осуществляли после обзора литературы, опираясь на консенсус экспертной группы, включавшей реаниматолога, врача общей практики, медсестры, прошедшей подготовку по интенсивной терапии, эпидемиологов и статистика. Перечень переменных был рассмотрен и одобрен 5 реаниматологами и биостатистиком, знакомых с базой данных ANZICS APD [283]. (Прогнозирование.)

«Для включения в наше предсказательное правило (*prediction rule*) из большего набора переменных мы выбрали 12 предикторов в соответствии с их клинической значимостью и результатами описательной статистики (*descriptive statistics*) исходных характеристик когорты пациентов отделения неотложной помощи с симптоматической фибрилляцией предсердий. В частности, учитывали исходные характеристики пациентов, у которых развилось и не развилось нежелательное явление (*adverse event*) в течение 30 суток, и отобрали для включения в модель 12 предикторов из 50 возможных в соответствии с очевидными различиями распределения значений предикторов между двумя группами, их клинической значимостью и чувствительностью (*sensibility*). <...> Для того чтобы ограничить коллинеарность и не нагружать модель (большим количеством переменных. — *Примеч. ред.*), рассчитали корреляцию по Спирмену между следующими клинически чувствительными ассоциациями: 1) гипертензия в анамнезе и прием β-блокаторов и диуретиков; 2) сердечная недостаточность в анамнезе и прием β-блокаторов и диуретиков в домашних условиях, периферический отёк при физической нагрузке и одышка в отделении неотложной помощи» [284]. (Прогнозирование.)

Пояснение

Часто исследователям доступно больше предикторов, чем они хотели бы включить в окончательную предсказательную модель. Поэтому требуется некоторая форма выбора предикторов. Для этой цели доступны различные методы, каждый из которых имеет свои сильные и слабые стороны (**рис. 2**).

Очевидный способ уменьшить количество потенциальных предикторов — изначально определить те, которые можно исключить (пункт 7а). Для этого можно изучить внешние доказательства, например, критически рассмотреть подходящую литературу, в идеале — в форме систематического обзора. Для сокращения числа потенциальных предикторов также важны и знания медицинских экспертов.

Предикторы могут исключать, исходя и из других соображений, например, по причине ненадёжности (*unreliable*) их измерения [58] или относительно высоких финансовых затрат или бремени, связанных с их измерением. Во втором случае иногда разрабатывают серию сложных моделей с такими предикторами и без них [262]. Кроме того, тесно связанные предикторы иногда могут объединять (например, методом статистического кластерного анализа или анализа главных компонент) в сводном показателе (например, наличие атеросклеротических симптомов [285]) или оценивать ассоциацию между предикторами (например, с использованием коэффициентов корреляции), чтобы предварительно выбрать 1 из 2 предикторов в случае их коллинеарности.

3. Выбор предикторов на этапе моделирования

Пример

«Мы использовали многофакторную логистическую регрессию с обратным пошаговым отбором и удаляли переменные при значении *P* больше 0,05; переменные (предикторы), которые, по нашему мнению, имеют большую клиническую значимость, были возвращены в модель. Мы также оценили дополнительные факторы риска (предикторы) из клинических руководств на предмет возможных дополнительных эффектов» [286]. (Диагностика.)

Пояснение

Даже если предварительно отобрать некоторые предикторы, как описано выше, все равно может остаться больше предикторов, чем хотелось бы включить в предсказательную модель (**рис. 2**). Последующий выбор предикторов может быть основан на прогностической значимости каждого из них или просто подогнанной модели с сохранением всех оставшихся предикторов [287].

Выбор предикторов может быть выполнен на основании силы их нескорректированной (одномерной) связи с предсказываемым исходом или путём их предварительной (перед многофакторным моделированием) селекции. Считается, что предикторы с ограниченной предсказательной ценностью по причине незначимой одномерной ассоциации с исходом могут быть исключены из процесса моделирования. Хотя эта стратегия довольно распространена, она не рекомендуется в качестве основы для выбора предикторов, поскольку важные предикторы могут быть отклонены из-за особенностей набора данных или искажений, вносимых другими предикторами [2, 112, 235]. Таким образом, незначимая (нескорректированная) статистическая ассоциация с изучаемым исходом не обязательно означает, что предиктор не важен. Однако если это сделано, следует сообщать о результатах однофакторного анализа, включая описание критериев выбора (например, уровень значимости) и размера выборки

(включая количество событий) для каждого предиктора (пункты 13б и 14б, **рис. 2**).

Распространённой процедурой в многофакторном моделировании является применение метода автоматического выбора переменных. В большинстве современных компьютерных программ доступно несколько вариантов, включая прямой отбор данных (*forward selection*), обратное исключение (*backward elimination*) и их комбинация. Обратное исключение начинается с полной модели (*full model*), включающей все потенциальные предикторы; переменные последовательно исключаются из модели до тех пор, пока не будет выполнено предварительно заданное правило остановки (такое как значение *P* или информационный критерий Akaike). Прямой отбор, напротив, начинается с пустой модели (*empty model*), в которую последовательно добавляют предикторы до тех пор, пока не будет выполнено предварительно заданное условие остановки.

При автоматическом выборе предикторов предпочтительнее использовать метод обратного исключения, учитывающий при моделировании все корреляции между предикторами [288]. Использование стратегий автоматического выбора предикторов для моделирования с несколькими переменными может привести к получению переобученных и оптимистичных моделей, особенно при небольшом размере выборки [2, 23–25, 32, 112, 289, 290]. Однако степень переобучения из-за использования стратегий выбора предикторов можно оценить и учесть в так называемых внутренних процедурах проверки (*internal validation procedures*) (**вставка B** и **рис. 1**).

Важным вопросом в автоматизированных процедурах выбора предикторов является критерий включения в модель [2]. Часто уровень значимости предиктора (α) устанавливается равным 0,05, как это принято при проверке гипотез. Однако симуляционные исследования показывают, что следует рассматривать более высокое значение, особенно в небольших наборах данных [25]. В таких случаях использование информационного критерия Akaike для выбора является привлекательным вариантом; критерий описывает согласованность модели данным, штрафуя за количество оцениваемых параметров и используя α=0,157 [2, 112, 291, 292].

Систематические обзоры многофакторных предсказательных моделей показали, что стратегия построения предсказательной модели часто остаётся неясной [34, 43, 54, 81, 182]. Например, из 11 исследований моделей прогнозирования аневризмального субарахноидального кровоизлияния в 36% подход к выбору предикторов был неясен [81].

4. Взаимодействия предикторов

Пример

«В модель были включены клинически значимые взаимодействия. Проверяли значимость всей группы взаимодействий, чтобы избежать увеличения ошибки типа I. Если результаты были незначимыми, всю группу взаимодействий исключали, а модель корректировали. В частности, изучали взаимодействие между применением β-блокаторов и диуретиков в домашних условиях, отёками, обнаруженными при физикальном обследовании, и сердечной недостаточностью в анамнезе» [284] (Прогнозирование).

Пояснение

Большинство предсказательных моделей включают предикторы в качестве основных эффектов, что предполагает, что эффекты всех предикторов являются аддитивными. Обратите внимание, что аддитивность (*additivity*) здесь предполагается в шкале моделирования: в единицах логарифма отношения шансов в случае логистической регрессии и логарифма отношения рисков для модели пропорциональных рисков Кокса (регрессионной модели). Аддитивность подразумевает мультипликативные эффекты в единицах оценки шансов и рисков соответственно [273]. Предположение об аддитивности означает, что предсказываемый эффект каждого предиктора не зависит от значений других предикторов. Это предположение может быть формально проверено путём оценки статистического взаимодействия между предикторами [112]. Немногие описания предсказательных моделей содержат сведения о взаимодействиях, и, похоже, мало кто из исследователей их изучает. Этот подход в целом разумен, поскольку эффекты взаимодействия редко повышают способность модели предсказывать изучаемые события.

Если изучить множество взаимодействий и включить в предсказательную модель только самые сильные, это приведёт к переобучению модели и в результате к чрезмерно оптимистичным оценкам её эффективности [2]. Авторы должны ограничить исследование взаимодействий небольшого количества предикторов с предварительным обоснованием этого списка, а не просто проверять все возможные взаимодействия, особенно при небольшом размере выборки. Альтернативой проверки взаимодействий является разработка разных моделей для разных подгрупп: например, для мужчин и женщин или взрослых и детей [278]. Однако из-за значительного сокращения размера выборки и соответствующей опасности переобучения модели этот подход используется редко, и его следует рассматривать только при большом размере выборки.

Модели прогнозирования выживаемости основываются на предположении, что эффекты предикторов постоянны во времени (т.е. риски пропорциональны) и они не взаимодействуют. Некоторые исследователи считают, что проверка гипотезы о пропорциональных рисках — это хорошая статистическая практика, тогда как другие предупреждают о рисках переобучения и оптимизма, если модели корректируются с учётом статистически значимых непропорциональных эффектов, аналогично тому, как описано выше для стратегий выбора предикторов [2, 112].

Авторы должны сообщить о процедурах проверки взаимодействий и пропорциональности рисков в моделях прогнозирования выживаемости, если таковая проводилась.

5. Внутренняя проверка

Пример

«Мы оценили внутреннюю достоверность (*internal validity*) с помощью процедуры бутстреппинга для получения реалистичных оценок эффективности обеих предсказательных моделей у будущих пациентов с похожими характеристиками. Воспроизвели весь процесс моделирования, включая выбор переменных… в 200 единицах наблюдения, отобранных с заменой из исходной выборки. Мы оценили эффективность выбранной предсказательной модели и простого правила, разработанных на основе каждой бутстреп-выборки, полученных из исходной выборки. Показатели эффективности в каждой бутстреп-выборке включали среднюю площадь под ROC-кривой, чувствительность и специфичность при оценке обоих исходов, а также уменьшение числа КТ-исследований при 100% чувствительности к нейрохирургическим вмешательствам» [286]. (Диагностика.)

Вставка Е. Внутренняя проверка

При разработке предсказательной модели можно получить чересчур оптимистичные оценки её эффективности. Этому могут способствовать следующие факторы: включение большого количества потенциальных предикторов по отношению к количеству событий исхода (при небольших размерах выборки), применение стратегий выбора предикторов (также при небольших размерах выборки) и категоризация непрерывных переменных [2, 12, 23–25, 32, 112, 290]. Задача исследователей — получить более достоверные оценки эффективности модели на основе данных, используемых для её разработки (*development data set*). Это можно сделать с помощью так называемой внутренней проверки (*internal validity*), предпочтительно с использованием методов повторного отбора данных, таких как бутстреппинг (*bootstrapping*), или перекрёстной проверки (*cross-validation*).

Предполагаемая эффективность

Предполагаемая эффективность (*apparent performance*) предсказательной модели оценивается непосредственно на основе набора данных, используемого для её разработки. Для небольших наборов данных это приводит к оптимистичным (смещённым, но стабильным) оценкам эффективности модели; однако при больших размерах выборки оптимистичность оценок снижается [32].

Проверка путём разделения выборки (разделение данных)

Классический вариант внутренней проверки — разделение набора данных, предназначенного для разработки модели (*development data set*), на две группы: один — для создания, другой — для проверки модели (см. **рис. 1** и **вставка В**). Обычно эти группы данных создают путём случайного разделения исходных данных (например, в соотношении 50 : 50 или 70 : 30). И, хотя, этот подход широко используется в исследованиях предсказательных моделей, у него есть ряд недостатков: 1) низкая эффективность (не используются все доступные для разработки данные); 2) оба набора данных будут очень похожи, потому что отличия между ними случайны (а значит, проверка модели, вероятно, покажет такой же результат, как и при использовании для разработки модели всего объёма данных); 3) разделение данных в разных соотношениях приведёт к разным результатам, особенно в небольших наборах данных [23, 25, 32, 295, 508]. Кроме того, неясно, сколько данных необходимо использовать для разработки модели и сколько отложить для её оценки (см. пункт 8). Применение этого подхода будет целесообразным при больших размерах выборки, и тогда полученная оценка эффективности модели обеспечит разумную оценку её реальной эффективности [2, 32]. Если размер выборки достаточно велик, лучшей альтернативой будет разделение набора данных по фактору времени (временна́я проверка, *temporal validation*) или местоположения (географическая проверка, *geographic validation*) [19, 20, 26].

Перекрёстная проверка

Перекрёстная проверка (*cross-validation*) — расширение метода разделения выборки; применяется с целью снижения систематической ошибки и изменчивости оценок эффективности модели [32]. Например, 10-кратная перекрёстная проверка предполагает случайное разделение данных на 10 групп одинакового размера. При этом модель разрабатывают на основе данных 9 групп, а эффективность оценивают по данным оставшейся группы. Эту процедуру повторяют 10 раз. Таким образом, данные каждой из 10 групп будут использованы для тестирования модели. Затем эффективность модели принимается как среднее значение 10 итераций.

Проверка методом бутстрепа

Метод бутстрепа не только позволяет использовать для проверки предсказательной модели все данные, но и обеспечивает механизм учёта переобучения модели или неопределённости, которые могут возникать в ходе всего процесса разработки, тем самым позволяя количественно оценить вероятность завышенных оценок эффективности окончательной предсказательной модели. Кроме того, этот метод позволяет получить оценку так называемого коэффициента сжатия (*shrinkage factor*), который может быть использован для корректировки коэффициентов регрессии и полученной оценки эффективности модели так, чтобы

в последующих исследованиях по проверке модели и её практическом применении можно было повысить её эффективность. Проверка методом бутстрепа включает [2, 12]:

1) разработку предсказательной модели с использованием данных всей исходной выборки (размером *n*) и определение предполагаемой эффективности (*apparent performance*);

2) создание бутстреп-выборки путём отбора *n* лиц с заменой из исходной выборки;

3) разработка модели на основе данных бутстреп-выборки (с применением тех же методов моделирования и отбора предикторов, что и на этапе 1):

а) определение предполагаемой эффективности модели (например, *c*-индекс), полученной на основе данных бутстреп-выборки (бутстреп-эффективность, *bootstrap performance*);

б) определение эффективности бутстреп-модели в исходной выборке (тестовая эффективность, *test performance*);

4) вычисление вероятности завышенной (оптимистичной) оценки как разницы между бутстреп- и тестовой эффективностью;

5) повторение шагов 2–4 не менее 100 раз;

6) усреднение оценки завышенных результатов, полученных на шаге 5, с вычитанием значения предполагаемой эффективности, полученной на шаге 1, чтобы получить оценку эффективности с поправкой на оптимизм (*optimism-corrected estimate of performance*).

Имеются доказательства того, что в многомерных исследованиях (например, омикс или полногеномные исследования ассоциаций) перекрёстная проверка (*cross-validation*) или бутстреппинг (*bootstrapping*) часто применяют некорректно из-за того, что авторы не повторяют все необходимые для моделирования шаги в каждой перекрёстной или бутсреп-выборке [299, 509, 510]. Это может привести к чрезмерно оптимистичной оценке эффективности модели [299, 511]. В совокупности этому могут способствовать и другие систематические ошибки [512].

Пояснение

Предиктивная эффективность модели на тех же данных, которые использовали для получения результатов предсказания, называется *предполагаемой эффективностью* (*apparent performance*) [12, 293, 294]. Это важно учитывать, так как многие предсказательные модели переобучены (*overfitted*), а их предполагаемая эффективность завышена (оптимистична), как правило, из-за применения стратегий выбора предикторов в небольших наборах данных [23–25, 32, 290, 295]. Более качественная начальная оценка эффективности предсказательной модели может быть достигнута путём использования методов повторной выборки (*resampling*), таких как перекрёстная проверка или бутстреппинг, называемых *внутренней проверкой* (*internal validation*) (**рис. 1** и **вставка E**) [12]. Мы рекомендуем, чтобы все исследования по разработке моделей включали какой-либо вариант внутренней проверки, особенно если не выполняется дополнительная внешняя проверка (*external validation*).

Выбор предикторов, основанный на предсказательной силе или значениях *P* в одно- и многофакторном анализе, часто приводит к значительной неопределённости в структуре модели [292, 296]. Преимущество метода бутстреппинга (в сравнении с перекрёстной проверкой) в качестве метода внутренней проверки заключается в том, что влияние стратегий выбора предикторов на построение модели и, соответственно, степень переобучения и оптимизма модели можно оценить количественно путём повторения процесса выбора в каждой бустреп-выборке [292, 296–298]. Кроме того, бутстреппинг обеспечивает оценку так называемого коэффициента поправки или коррекции, с помощью которого модель (т.е. коэффициенты регрессии) и показатели её эффективности (пункт 16) могут быть уменьшены, и, таким образом, скорректировано переобучение (**вставка E**). Крайне важно, чтобы все аспекты подбора модели были учтены в каждой случайной или бутстреп-выборке, включая выбор предикторов, решения о преобразовании данных и проверку взаимодействия с другими переменными или со временем. Пропуск этих шагов является обычным в клинических исследованиях, что может привести к смещённым оценкам согласованности даже в проверочной выборке (*validation sample*) [299, 300]. Не рекомендуется повторно выбирать одни и те же предикторы в каждой бустреп-выборке, если только модель не была построена с использованием всех предикторов (так называемая полная модель, *full model*). Авторы должны подробно описать все процедуры, выполняемые при внутренней проверке модели.

Переобучение (*overfitting*), завышенные оценки (*optimism*), ошибки калибровки модели (*miscalibration*) можно также решить путём применения процедур сжатия (*shrinkage*) или штрафа [287, 290, 294, 301]. В случае разработки модели событий или модели с большим количеством предикторов при небольшом размере выборки особенно популярен метод лассо (*lasso method*) и его варианты [24, 302, 303]. Однако его полезность при меньшем количестве предикторов менее очевидна [291]. Если такая процедура была проведена, следует подробно описать использованный метод [например, метод лассо, гребневая регрессия (*ridge regression*), эвристическое сжатие (*heuristic shrinkage*)].

Обзор исследований предсказательных моделей показал, что только в 5 из 14 работ, опубликованных в медицинских журналах общего профиля, сообщалось о внутренней проверке [34]; аналогичные результаты показаны и в других обзорах [43, 53, 55, 64, 66, 71, 75, 76, 88, 93–95, 304, 305].

Пункт 10в. Для проверочных исследований опишите, как рассчитывали вероятности предсказываемого исхода (П).

Примеры

«Эффективность вычисления риска рака предстательной железы определяли путём получения предсказываемой вероятности любого случая рака простаты, а также агрессивного рака простаты для каждого пациента с помощью PRC (калькулятора риска Prostate Cancer Prevention Trial. — *Примеч. авт.*) (http://deb.uthscsa.edu/URORiskCalc/Pages/uroriskcalc.jsp) и SRC (Sunnybrook nomogram-based prostate cancer risk calculator. — *Примеч. авт.*) (www.prostaterisk.ca)» [306]. (Диагностика.)

«Индекс HSI (Hepatic Steatosis Index. — *Примеч. авт.*) для определения вероятности развития стеатоза печени рассчитывали по формуле Lee и соавт. [ссылка]:

$$HSI = \frac{e^{0{,}315 \times \text{ИМТ} + 2{,}421 \times \text{АЛТ/АСТ} + 0{,}630 \times \text{СД} - 9{,}960}}{1 + e^{0{,}315 \times \text{ИМТ} + 2{,}421 \times \text{АЛТ/АСТ} + 0{,}630 \times \text{СД} - 9{,}960}},$$

где при наличии сахарного диабета (СД) указывали 1, при отсутствии — 0; АЛТ — аланинаминотрансфераза, АСТ — аспартатаминотрансфераза» [307]. (Диагностика.)

«Открытый код для расчёта риска развития колоректального рака с помощью QCancer доступен по адресу www.qcancer.org/colorectal/ (распространяется по лицензии GNU Lesser General Public Licence, версия 3)» [308]. (Прогнозирование.)

Пояснение

Эффективность существующей предсказательной модели для нового набора данных (**вставка В** и **рис. 1**) предпочтительно оценивать, основываясь на предсказании, выполненном с помощью оригинальной модели (как опубликовано), путём сравнения этих предсказаний с актуальными исходами в наборе проверочных данных (и, таким образом, проводить калибровку и проверку способности модели различать исходы) [309] (пункт 10г). Поэтому важно, чтобы авторы, оценившие эффективность существующей предсказательной модели, ясно и однозначно указали, как они получили соответствующие предсказания. Такое описание может включать представление полной модели [все коэффициенты регрессии (*regression coefficients*), включая свободный коэффициент (*intercept*) или исходные риски (*baseline hazard*) в определённый момент времени], ссылки на веб-калькулятор или компьютерный код для реализации модели предсказания (пункт 14).

В некоторых исследованиях по разработке моделей бывает представлено несколько моделей или вариантов одной модели (например, как полная регрессионная модель, так и упрощённая шкала). При необходимости авторы должны уточнить, какую именно из этих предсказательных моделей они оценивали.

Предсказательные модели часто представляют графически в виде номограмм (пункт 15б) [310, 311], которые позволяют выполнять расчёты для отдельных лиц без калькулятора или компьютера. Но в проверочных исследованиях с большим числом участников они неприменимы. Авторы должны ясно и однозначно объяснить, как были получены предсказания — вручную с использованием фактической номограммы или с применением лежащей в её основе регрессионной модели.

Без доступа к опубликованной предсказательной модели последующая проверка, повторная калибровка и обновление невозможны. Например, модель FRAX для прогнозирования 10-летнего риска остеопороза или перелома бедра [312], которая в настоящее время включена в многочисленные клинические руководства по всему миру [35, 37, 313], не была опубликована, что делает независимую оценку модели невозможной [314–316].

Существует ряд ошибочных представлений о том, как проверить существующую модель. Одно из них заключается в желании повторить весь процесс моделирования на основе новых данных, включая выбор предиктора и оценку коэффициентов регрессии (и эффективности модели), а затем сравнить полученные результаты с оригинальными. Другое заблуждение — переделать ранее разработанную и опубликованную модель в окончательную модель с использованием проверочных данных. В обоих случаях результатом будет фактически другая, новая модель, а не проверка существующей [19, 20, 26, 28, 47, 309].

Авторы часто проверяют разработанную ими предсказательную модель на другом наборе данных (например, набранных позже или в другом медицинском учреждении). Если оценки эффективности модели, полученные на основе данных для разработки и проверки модели, признаются одинаковыми, нередки случаи последующего объединения обоих наборов данных и разработки на их основе новой предсказательной модели [317]. Хотя само по себе это не плохо, но такое исследование не является проверочным, а скорее представляет собой проверку и разработку модели или её повторную разработку. Полученную таким образом модель необходимо в дальнейшем снова проверять.

Пункт 10г. Укажите все показатели, с помощью которых оценивали эффективность модели и, если применимо, сравнивали несколько моделей (Р; П).

Существует множество показателей для определения и количественной оценки предсказательной эффективности соответствующих моделей [26, 252, 253, 269] (**вставки Ж** и **З**). Здесь мы выскажемся в защиту наиболее широко используемых показателей, о которых мы рекомендуем сообщать исследователям. Это традиционные (статистические) показатели, а также показатели, предложенные сравнительно недавно, которые в той или иной степени учитывают клинические последствия предсказаний и показатели для оценки дополнительной предсказательной ценности (*incremental predictive value*) конкретного предиктора, помимо существующих

или установленных предикторов, или при сравнении различных моделей.

1. Традиционные показатели

Примеры

«Мы оценили предсказательную эффективность шкалы риска QRISK2-2011 в когорте THIN (The Health Improvement Network. — *Примеч. ред.*) путём калибровки и дискриминации. Калибровка позволяет оценить согласованность предсказываемого и наблюдаемого 10-летнего сердечно-сосудистого риска. Согласованность была определена путём вычисления отношения прогнозируемого сердечно-сосудистого риска к наблюдаемому отдельно для мужчин и женщин для каждого дециля (10-й доли) предсказанного риска, обеспечив 10 групп одинакового размера и каждого пятилетнего возрастного диапазона. Калибровку предсказаний оценки риска выполняли путём построения графика наблюдаемых долей против предсказываемых вероятностей с вычислением углового коэффициента (*calibration slope*).

«Дискриминация — это способность оценки риска различать пациентов, у которых наступило и не наступило предсказываемое событие на протяжении периода исследования. Этот показатель определяется количественно путём вычисления площади под ROC-кривой, где значение 0,5 указывает на случайную, а 1 — идеальную дискриминацию. Также рассчитали коэффициенты D и R^2 — показатели дискриминации и объяснённой дисперсии соответственно, адаптированные к цензурированным данным о выживаемости. Более высокие значения коэффициента D указывают на более точную дискриминацию, где увеличение значений коэффициента на 0,1, по сравнению с другими оценками риска, является хорошим признаком повышения прогностической точности» [117]. (Прогнозирование, Проверка.)

«Во-первых, используя ROC-анализ, мы сравнили возможности клинического правила принятия решения (*clinical decision rule*) и суждения врача общей практики в различении пациентов с заболеванием и без него. Площадь под ROC-кривой (AUC — *area under the ROC curve*) со значением 0,5 указывает на отсутствие дискриминации, со значением 1,0 — на идеальную дискриминацию. Затем мы построили калибровочный график, чтобы отдельно изучить соответствие между предсказанными вероятностями правила принятия решения и наблюдаемым исходом острого коронарного синдрома, и аналогичный калибровочный график для предсказанных вероятностей врача общей практики. Предсказания, идеально согласующиеся с наблюдаемым исходом, должны лежать на калибровочной линии под углом 45°» [318]. (Диагностика, Разработка.)

«Точность проверенной на исходных данных и скорректированной модели проверяли на новом наборе проверочных данных. Формулу регрессии из разработанной модели мы применяли ко всем наблюдениям (работники пекарни) из этого набора данных. С целью калибровки оценивали согласованность между предсказываемыми вероятностями и наблюдаемой частотой путём построения графика зависимости вероятностей (ось *x*) от наблюдаемой частоты исхода (ось *y*). Ассоциация между предсказанными вероятностями и наблюдаемыми частотами может быть описана линией, задаваемой свободным (*intercept*) и угловым коэффициентами (*slope*). Идеальная калибровка считается достигнутой при значении свободного коэффициента, равным 0, и углового коэффициента, равным 1. <...> Качество дискриминации оценивали по величине площади под ROC-кривой» [319]. (Диагностика, Разработка.)

Пояснение

Два ключевых аспекта характеризуют эффективность предсказательной модели: калибровка (*calibration*) и дискриминация (*discrimination*). Их следует описывать во всех статьях о предсказательных моделях (**вставки Ж** и **З**).

Калибровка отражает согласованность между предсказываемыми вероятностями и наблюдаемыми исходами. Калибровку предпочтительно представлять в графическом виде, где наблюдаемые риски откладывают по оси *y,* прогнозируемые риски — по оси *x*, но также эту информацию можно представить в виде таблицы.

Дискриминация — способность предсказательной модели различать пациентов с целевым исходом или без него. Наиболее общий и часто используемый показатель дискриминации (в логистических моделях и моделях выживаемости) — это индекс согласованности (*c*-индекс), который равен площади под ROC-кривой для логистических предсказательных моделей. Существует несколько различных вариантов *c*-индекса [320], поэтому авторы должны указать, какой именно вариант они рассчитывали.

В дополнение к показателям дискриминации и калибровки могут быть представлены и другие показатели *общей эффективности* (*overall performance*) модели. К ним относятся показатели объяснённой дисперсии (*explained variation*, R^2) [321–329] и оценки Бриера (*Brier score*) [330–332]. Предложено множество различных подходов к вычислению величины R^2, поэтому авторам следует чётко указать, какой из них они использовали в своей работе. Для моделей выживания (например, разработанных методом регрессии Кокса) недавно был предложен коэффициент D в качестве показателя прогностической диференциации [333].

Вставка Ж. Показатели эффективности

После того как мы разработаем предсказательную модель, мы должны оценить её эффективность. Наиболее важными аспектами эффективности модели являются дискриминация и калибровка (**пункт 10г** и **вставка З**). В исследованиях по разработке моделей (*model development studies*) нас в первую очередь интересует дискриминация, поскольку модель будет

хорошо откалибрована (в среднем) по определению. В проверочных исследованиях (*validation studies*) оценка как дискриминации, так и калибровки имеет фундаментальное значение [252, 513].

Калибровка (*calibration*) отражает согласованность между предсказаниями исхода, полученными моделью, и наблюдаемыми исходами. Неформально модель считают хорошо откалиброванной, если каждая группа, например, из 100 человек со средним предсказанным риском *x%* имеет (диагностическая модель) или развивает (прогностическая модель) предсказываемый исход с частотой, близкой к *x*.

Калибровку предпочтительно отображать в графическом виде, откладывая предсказываемые вероятности исхода на оси *x*, а наблюдаемую частоту исхода — на оси *y*. Этот график обычно строится с помощью десятых долей предсказываемого риска и, что предпочтительно, дополняется сглаженной (*lowess*) линией по всему диапазону предсказываемых вероятностей, что возможно для предсказательных моделей, разработанных как методом логистической регрессии [112, 514], так и с помощью моделирования выживаемости [515] (см. пункт 16). Такой график отображает направление и величину ошибок калибровки модели (*miscalibration*) по всему диапазону вероятностей, которые можно комбинировать с оценками калибровочных углового (*slope*) и свободного (*intercept*) коэффициентов [515]. Для хорошо откалиброванной модели наиболее точные предсказания (при сглаживании или использовании подгрупп) будут находиться на линии калибровочного графика, идущей под углом около 45°. Идеальная калибровка показывает угловой коэффициент, равный 1, и свободный коэффициент, равный 0, но с некоторыми оговорками [516].

Калибровочные графики, как правило, показывают хорошую калибровку в наборе данных, на основе которых модель была разработана, и даже идеальную, когда используется метод сглаживания. Дополнительно можно провести тест на соответствие калибровочных свободного и углового коэффициентов 0 и 1 соответственно [517, 518]. Сравнение вероятностей предсказываемых и наблюдаемых исходов можно представить также в таблице (обычно для десятых долей предсказываемого риска).

Наконец, для статистической проверки согласованности между предсказанными и наблюдаемыми вероятностями широко применяют тест Хосмера–Лемешова (*Hosmer-Lemeshow test*) или аналогичные для моделей выживаемости, включая тест Нама–Д'Агостино [519] или Гроннесби–Боргана (*Grønnesby-Borgan test*) [520]. Такие тесты обладают ограниченной статистической мощностью для оценки плохой калибровки и чувствительны к группировке данных и размеру выборки [521–523], часто они незначимы для малых *N* и почти всегда значимы для больших *N*. Кроме того, они не отражают направление и величину ошибок калибровки, поэтому предпочтение отдаётся калибровочным графикам.

Кроме того, калибровку (графически) можно также проводить в подгруппах, сформированных по признакам ключевых предикторов, в частности, по возрасту или полу [117, 524]. Недавно были предложены методы оценки калибровки полиномиальных предсказательных моделей [525].

Дискриминация (*discrimination*) означает способность предсказательной модели различать тех, у кого наступает или не наступает событие исхода. Дискриминация считается идеальной в том случае, если предсказываемые риски у всех лиц, у которых событие исхода имелось (диагностика) или развилось (прогнозирование), будут выше, чем у всех лиц, у которых такого события не было. Дискриминацию обычно оценивают по так называемому индексу согласованности (*c-index*). Этот индекс отражает вероятность того, что для любой случайно выбранной пары индивидуумов (с целевым состоянием и без него) модель присваивает более высокую вероятность индивидууму с целевым состоянием [526]. *C*-индекс идентичен площади под ROC-кривой для моделей с бинарными конечными точками (*endpoints*) и может быть обобщён для моделей выживаемости (время до события, *time-to-event*) с учётом цензурирования. Для последних был предложен ряд различных *c*-индексов [527], поэтому авторы должны чётко указать, какой показатель используется, и привести соответствующую ссылку. Кроме того, недавно предложены расширения *c*-индекса для моделей с более чем двумя категориями исходов [528], конкурирующими рисками [529] и кластеризацией [170, 171].

Показатели общей эффективности (*overall performance measures*), такие как объяснённая дисперсия (*explained variation*, R^2) [321, 324–329] и оценка Бриера (*Brier score*) [330, 331], иногда указывают в дополнение к традиционным показателям дискриминации и калибровки, хотя их интерпретация менее очевидна. Предложено много различных подходов к вычислению величины R^2, поэтому важно, чтобы авторы чётко определяли версию, которую они вычисляли и о которой сообщают в своей работе.

Классификационные показатели (*classification measures*), такие как предсказательная ценность (*predictive values*), чувствительность (*sensitivity*) и специфичность (*specificity*), являются показателями эффективности после введения одного или более пороговых значений вероятности. С их помощью можно оценить точность (*accuracy*) или показатели классификации, часто сообщаемые в отдельных исследованиях диагностических тестов или прогностических факторов. Однако такая дихотомизация и связанные с ней показатели классификации приводят к потере информации. Более того, введение такого порога подразумевает,

что он актуален в клинической практике, что бывает не так часто.

Анализ кривой принятия решений (*decision curve analysis*) [360, 363–366] даёт представление о клинических последствиях путём определения зависимости между выбранным предсказываемым порогом вероятности и относительным значением ложноположительных и ложноотрицательных результатов для получения оценки пользы (*net benefit*) от применения модели при таком пороговом значении.

Индекс реклассификации (*net reclassification improvement,* NRI) обычно используется для количественной оценки пользы добавления нового предиктора к существующей модели или для сравнения двух невложенных (*nonnested*) моделей [339, 347, 348, 420, 530]. NRI — это сумма долей правильно реклассифицированных наблюдений с наступившим и ненаступившим исходом. Верхняя граница NRI — это непрерывный NRI (т.е. без категорий), который учитывает любое изменение (увеличение или уменьшение) предсказываемого риска для каждого индивидуума [347, 530].

Интегрированный индекс дискриминации (*integrated discrimination improvement,* IDI) — это разница в предсказываемых вероятностях между теми, у кого имеется/отсутствует (диагностика) или произошло / не произошло развитие (прогнозирование) предсказываемого события [339]. По этой разнице оценивается изменение (повышение или снижение) разницы вероятности исхода между двумя моделями (вложенными или не вложенными) по всем возможным порогам вероятности. Индекс можно интерпретировать как эквивалент разницы средней предсказываемой вероятности у лиц без исхода и с ним.

Вставка 3. Оценка эффективности регрессионной модели Кокса

Оценить эффективность большинства регрессионных предсказательных моделей несложно. Общепринятый подход для моделей логистической регрессии — построение графика наблюдаемой вероятности исхода против предсказываемой вероятности для нескольких групп (обычно 10), определяемых предсказываемым риском (см. **вставку Ж** и пункты 10г и 15а). На таком калибровочном графике способность модели различать наблюдения (*discrimination*) определяется разбросом предсказанных вероятностей по группам риска [417]; также могут быть получены и формальные показатели дискриминации (пункт 10г).

Похожий подход может применяться к полностью параметрическим моделям (*fully parametric models*) для данных типа «время до события» (*time-to-event*), но эти модели используются редко. Для таких данных предпочтительно использовать регрессию Кокса (*Cox regression*), но калибровку модели Кокса провести сложнее, поскольку модель Кокса не полностью определена. Модель позволяет оценить относительные различия риска между пациентами с разными характеристиками, но, она не оценивает исходную (без предикторов. – *Примеч. ред.*) функцию выживания (*baseline survival function*), а значит не оценивает и абсолютные риски (вероятности событий) [309]. Исключением является случай, когда целью предсказательной модели, основанной на регрессии Кокса, являются исходы в фиксированный момент времени (например, риск смерти от сердечно-сосудистых заболеваний в течение 10 лет). В этом случае требуется только исходная вероятность выживания в интересующий момент времени, а дискриминацию и калибровку можно провести с помощью методов, описанных во **вставке Ж**.

Построение модели Кокса

Модель Кокса строится на основе набора предикторов, выраженных соответствующими коэффициентами регрессии (логарифмированные значения отношения рисков, *log hazard ratios*) [411, 531]. Прогностический индекс (*prognostic index*) — это взвешенная сумма переменных в модели, где веса являются коэффициентами регрессии (см. пример в пункте 15б). Прогностический индекс для индивидуума представляет собой логарифм относительного риска по сравнению с гипотетическим индивидуумом, чей индекс равен нулю [309].

Когда новая модель получена с использованием регрессии Кокса, прогностический индекс может быть использован для изучения предсказываемой выживаемости для нескольких групп риска. Например, пациенты могут быть разделены на 4 равные группы исходя из значений индекса. В этом случае дискриминацию можно оценить как визуально по разбросу кривых Каплана–Мейера для этих групп риска, так и путём получения численных показателей эффективности модели (см. пункт 10г). Калибровка может быть изучена путём наложения кривых выживаемости, полученных непосредственно из модели Кокса [309, 373].

Проверка модели Кокса

На практике исходную функцию выживаемости для модели Кокса никогда не публикуют. В результате внешняя проверка модели Кокса разными исследователями затруднена, поскольку невозможно оценить абсолютные риски, особенно нелегко выполнить калибровку. Royston и Altman [309] предложили различные варианты анализа в зависимости от объёма информации, полученного из исследования, в котором была получена соответствующая модель (*derivation study*).

Способность модели различать наблюдения с разным исходом (*discrimination*) можно определить при условии, что проверяемая модель (*derivation model*) представлена как минимум набором предикторов с определёнными коэффициентами регрессии и точно известно кодирование каждого предиктора. Затем для каждого участника в проверочном наборе данных

(*validation data set*) может быть вычислено значение прогностического индекса, а в последующем выполнен регрессионный анализ с использованием значений индекса в качестве единственной переменной (*covariate*). При похожем наборе наблюдений, дискриминация в наборе проверочных данных примерно такая же, как и в исходных данных (*derivation data*), а коэффициент регрессии для прогностического индекса будет равен приблизительно единице. Если угловой коэффициент (*slope*) в проверочном наборе <1, дискриминация будет хуже, если >1 — лучше.

Если для нескольких групп риска в исходном исследовании (*derivation study*) представлены графики Каплана–Мейера, то сравнение соответствующих графиков для исходного и проверочного наборов данных даёт лишь грубое представление о калибровке модели. О хорошей калибровке можно говорить (на основе суждения, а не формального сравнения), если хорошо согласуются два набора кривых выживаемости. Однако такая оценка калибровки не является примером строгого сравнения между наблюдаемыми и предсказываемыми значениями, поскольку модель Кокса не используется напрямую для предсказывания вероятностей выживания. Без исходной функции выживания невозможно судить о том, насколько хорошие результаты покажет калибровка в независимой выборке [309].

Для сравнения эффективности разных моделей на одном и том же наборе данных можно использовать формальные статистические тесты, такие как тест DeLong [334]. Однако этот тест неприменим, если сравниваемые модели являются вложенными (*nested*) (т.е. если одна модель содержит все предикторы другой и как минимум один дополнительный предиктор) и получены на основе одного и того же набора данных, например, если модель с клиническими предикторами и новым молекулярным маркером сравнивают с моделью, в которой присутствуют только клинические предикторы [335].

И наконец, любую вновь разработанную модель рекомендуется широко сравнивать с существующими опубликованными моделями, в идеале по количественным показателям [47, 48]. В отсутствие какого-либо прямого сравнения между двумя или более моделями на одном и том же наборе данных сложно выбрать из всех доступных предсказательных моделей потенциально более полезную. Многочисленные систематические обзоры показали, что в немногих исследованиях по разработке или проверке предсказательных моделей для одного и того же исхода сравнивают их эффективность с другими существующими моделями [82].

Калибровка (*calibration*) — ключевая характеристика, и её проведение широко рекомендуется. Однако многие систематические обзоры многофакторных предсказательных моделей показали, что о калибровке редко сообщают [34, 41, 43, 55, 62, 63, 66, 73–82, 84, 86, 88, 90–92, 94, 122, 176, 180, 267, 336]. Например, калибровку проводили только в 10 из 39 (26%) исследований моделей предсказания диабета 2-го типа [45]. Дискриминация (*discrimination*) — наиболее часто измеряемый показатель эффективности, но о ней тем не менее также не всегда сообщают [74, 78, 81, 88, 122, 336, 337] (например, лишь в 44% случаях описания моделей аневризматического субарахноидального кровоизлияния [81]). Очень немногие исследования включают сравнение эффективности модели с другими существующими предсказательными моделями с использованием одного набора данных [81, 82, 122].

2. Количественная оценка ценности дополнительного предиктора

Пример

«Оценили дополнительную прогностическую ценность (*incremental prognostic value*) биомаркеров при добавлении к шкале GRACE с помощью теста отношения правдоподобия (*likelihood ratio test*). Для оценки величины приращения эффективности модели при добавлении отдельных биомаркеров к шкале GRACE использовали три дополнительных показателя дискриминации: изменение площади под ROC-кривой (ΔAUC), интегрированный индекс дискриминации (*integrated discrimination improvement*, IDI) и индекс реклассификации (*net reclassification improvement*, NRI) в непрерывной и категориальной шкале измерения. Для определения клинической полезности рассчитали NRI (>0,02), в соответствии с которым 2% рассматривается как минимальный порог значимого изменения прогнозируемого риска. Кроме того, добавили два категориальных NRI с заранее определёнными порогами риска 6 и 14%, выбранными в соответствии с предыдущим исследованием, или 5 и 12% согласно наблюдаемой частоте событий в настоящем исследовании. Категориальные NRI определяют повышение или понижение частоты реклассификации только в том случае, если предсказываемые риски переходят из одной категории в другую. Поскольку количество биомаркеров, добавленных к шкале GRACE, оставалось небольшим (максимум 2), степень чрезмерного оптимизма, вероятно, была незначительной. Тем не менее на этапе внутренней проверки повторно рассчитали показатели ΔAUC и IDI методом бутстрепа и подтвердили полученные результаты» [338]. (Прогнозирование; Дополнительная ценность.)

Пояснение

Преимущество многофакторного анализа, по сравнению с исследованиями с одним маркером или тестом, состоит в том, что он позволяет определить дополнительную ценность (*incremental value*) теста или маркера. Однако количественную оценку дополнительной ценности определённого, зачастую нового предиктора, добавленного к уже известным предикторам или даже к существующей предсказательной модели, на основании увеличения

или улучшения в целом традиционных показателей эффективности (калибровка, дискриминация или R^2), трудно интерпретировать клинически [339, 340]. Кроме того, есть опасения, что такие показатели эффективности, как *c*-индекс, нечувствительны для оценки дополнительной ценности [341, 342], хотя его роль в качестве описательного показателя по-прежнему остаётся полезной [343]. Наконец, тесты на статистическую значимость могут ввести в заблуждение, потому что статистически значимые ассоциации новых, но слабых предикторов легко обнаружить в большой выборке.

По этой причине были предложены новые показатели, основанные на концепции повторной классификации (*reclassification*) лиц по заранее определённым категориям риска. Представление этой информации в табличном виде покажет распределение отдельных лиц по новым категориям риска (от низкого к высокому и наоборот), определённым с помощью модели с конкретным предиктором или без него [344–346]. Использование таблиц реклассификации явно зависит от выбора порогов для определения групп риска (пункт 11).

Индекс реклассификации (*net reclassification improvement,* NRI) — широко используемый показатель для количественной оценки частоты реклассификации, наблюдаемой в таблицах [339, 347, 348]. NRI можно использовать при разработке модели в случае добавления определённого предиктора к установленным предикторам или существующей модели (т.е. модели являются вложенными, *nested*), а также при проверке модели путём сравнения невложенных (*nonnested*) моделей при условии, что сравниваемые модели достаточно хорошо откалиброваны [349]. Следовательно, перед использованием NRI сначала необходимо оценить калибровку модели, чтобы читатели могли судить об уместности вычисления NRI.

Показано, что NRI чрезвычайно чувствителен к выбору пороговых значений, определяющих категории риска (и, таким образом, открыт для манипуляций), и, кроме того, существует несколько других предостережений относительно его использования, особенно в моделях с субоптимальной калибровкой [350–356]. Поэтому мы рекомендуем расчёты NRI всегда сопровождать таблицей классификации, стратифицированной для участников с целевым исходом и без него [357] (пункт 16). Также высказывались опасения, что непрерывный NRI, который является мерой связи, а не улучшения модели, может быть причиной ошибочной интерпретации и чувствителен к неправильной калибровке модели (*model miscalibration*) [346].

По сравнению NRI, предпочтительнее использовать такие показатели, как изменение чистой выгоды (*change in net benefit*), изменение относительной полезности (*change in relative utility*) и взвешенный индекс реклассификации (*weighted net reclassification improvement*). Эти три показателя могут быть математически преобразованы друг в друга [349]. Определение подходящих показателей для количественной оценки дополнительной ценности добавления предиктора к существующей предсказательной модели остаётся областью активных исследований. Также сохраняет привлекательность поиск клинически интуитивных показателей с помощью основанного на модели теста отношения правдоподобия *(likelihood ratio*) [343, 358].

Систематические обзоры показали, что авторы исследований по реклассификации данных редко сообщают о том, чем был обусловлен выбор пороговых значений рисков [105]. Кроме того, в половине исследований не сообщалось о калибровке моделей, и лишь немногие представляли результаты правильной и ошибочной реклассификации данных.

3. Показатели полезности

Пример

«Мы использовали анализ кривой принятия решений (*decision curve analysis*) (с учётом цензурированных наблюдений) для описания и сравнения клинических эффектов шкалы риска QRISK2-2011 и уравнения NICE Framingham. Считали, что модель имеет клиническую ценность, если она обеспечивает наибольшую чистую выгоду (*net benefit*) в диапазоне пороговых значений, при которых индивидуум может быть отнесён к группе высокого риска. Вкратце чистая выгода модели — это разница долей истинно положительных и ложноположительных результатов, последние — взвешенные на величину выбранного порогового значения для обозначения высокого риска. При любом заданном пороговом значении предпочтительной будет модель с более высокими показателями чистой выгоды» [117]. (Прогнозирование; Проверка.)

Пояснение

Дискриминация и калибровка — статистические характеристики эффективности предсказательной модели. Однако клинические последствия принятия конкретного уровня дискриминации или ошибочной калибровки сложно предвидеть [359, 360]. Для получения представления о клинических последствиях или чистой выгоде (*net benefit*) от использования предсказательной модели при определённых пороговых значениях [349] предложены новые подходы, такие как анализ кривой принятия решений (*decision curve analysis*) [361–363] и оценка относительной полезности (*relative utility*) [364–366]. Их также можно использовать для сравнения клинической полезности различных моделей: например, первоначальной и расширенной моделей, протестированных на одном и том же наборе данных, или даже двух разных моделей (разработанных на двух разных наборах данных), проверенных на одном и том же независимом наборе данных [367].

Пункт 10д. Опишите любое обновление модели (например, повторную калибровку), выполненное в результате её проверки (если применимо) (П).

Примеры

«Коэффициенты (оригинальной диагностической. — *Примеч. авт.*) экспертной модели, вероятно, подвержены

переобучению, поскольку исходно рассматривали 25 диагностических показателей и только 36 эпизодов. Чтобы количественно оценить переобучение, мы определили (в нашем наборе проверочных данных. — *Примеч. авт.*) коэффициент сжатия (*shrinkage factor*), изучая наклон калибровки *b* при подборе модели логистической регрессии:

$$\text{logit}\,(P\,(Y=1))=a + b * \text{logit}\,(p),$$

где [Y=1 указывает на наличие пневмонии (исход) в проверочном наборе данных. — *Примеч. авт.*] p — вектор предсказываемых вероятностей. Угол наклона *b* линейного предиктора определяет коэффициент сжатия. В хорошо откалиброванных моделях величина *b* приблизительно равна 1. Таким образом, мы повторно калибруем коэффициенты настоящей экспертной модели, умножая их на коэффициент сжатия (сжатие после оценки)» [368]. (Диагностика; Обновление модели; Логистический.)

«В этом исследовании мы применили метод (обновления модели. — *Примеч. авт.*) проверки путём калибровки, предложенный Van Houwelingen. Для каждой категории риска была подобрана модель пропорциональных рисков Вейбулла (*Weibull*) с использованием значений общей выживаемости, предсказанных (оригинальной. — *Примеч. авт.*) моделью UISS. Эти ожидаемые кривые были сопоставлены с наблюдаемыми кривыми Каплана–Мейера, а возможные различия оценивали с помощью калибровочной модели, которая определяла, насколько оригинальная прогностическая оценка была воспроизводима на новых данных путём тестирования трёх различных параметров — α, β и γ. Если нулевая гипотеза при α=0, β=–1 и γ=1 была отклонена (т.е. если обнаруживались расхождения между наблюдаемыми и ожидаемыми кривыми), оценки калибровочной модели использовали для повторной калибровки предсказываемых вероятностей. Отметим, что повторная калибровка не влияет на точность дискриминации данных моделью. Конкретные детали этого подхода описаны в статьях Van Houwelingen и Miceli и соавт.» [369]. (Прогнозирование; Обновление модели; Выживаемость.)

«Результаты внешней проверки побудили нас обновить модели. Мы скорректировали свободный коэффициент (*intercept*) и коэффициенты регрессии предсказательных моделей для ирландской выборки. Наиболее важное отличие от результатов, полученных на голландской выборке, — более низкий пороговый уровень гемоглобина, приемлемый для донорства, который влияет на исход и точку изменения (*breakpoint*) в кусочно-линейной функции (*piecewise linear function*) для предикторов предыдущего уровня гемоглобина. Для обновления применяли два метода: повторную калибровку модели и пересмотр модели (*model revision*). Повторная калибровка включала корректировку свободного коэффициента и отдельных коэффициентов регрессии с одинаковым угловым коэффициентом калибровки (*calibration slope*). Для пересмотренных моделей каждый коэффициент регрессии был скорректирован отдельно путём пошагового добавления предикторов к повторно откалиброванной модели и проверки их дополнительной ценности с помощью критерия отношения правдоподобия (p <0,05). Если последнее подтверждалось, коэффициент регрессии для предиктора корректировали» [370]. (Диагностика; Обновление модели; Логистический.)

Пояснение

При проверке (или применении) существующей предсказательной модели на других людях предсказательная эффективность обычно ниже, чем на данных, которые были использованы для разработки модели. Причем разница в эффективности тем больше, чем более строгая форма проверки применяется (**вставка B** и **рис. 1**). Снижение эффективности более вероятно при проверке в других географических условиях (*geographic validation*) или условиях наблюдения (*setting validation*), а также другими исследователями, в сравнении с результатами предварительной проверки (*temporal validation*) одними и теми же исследователями [2, 20, 21, 102, 290]. При относительно низкой точности предсказания (*predictive accuracy*) исследователи могут отказаться от существующей модели, скорректировать её в своем проверочном наборе данных (*validation set*) или даже разработать совершенно новую модель.

Разработка новой модели для предсказания тех же исходов или в той же целевой популяции — заманчивый выход из ситуации, но по разным причинам такой подход нежелателен [20, 31, 102, 290]. Во-первых, разработка разных моделей для разных периодов времени, стационаров, стран или условий наблюдения ограничивает применимость результатов предсказательных исследований в иных условиях. Во-вторых, медицинским организациям и работникам будет непросто выбрать подходящую модель из множества ей подобных. В-третьих, в проверочные исследования часто включают меньше людей, чем в исследования по разработке соответствующей модели, что делает новую модель более подверженной переобучению (*overfitting*) и, возможно, даже менее обобщаемой (*generalizable*), чем оригинальная модель. Наконец, предшествующие знания, полученные в оригинальных (по разработке модели) исследованиях, не используются оптимально, что противоречит представлению о том, что выводы и рекомендации по совершенствованию доказательной медицины должны основываться на как можно большем количестве данных [371].

Прежде чем разработать новую модель на основе имеющихся проверочных данных, можно сначала попытаться скорректировать (т.е. обновить) оригинальную предсказательную модель, чтобы определить, в какой степени потеря точности предсказания может быть преодолена [85]. Преимущество скорректированной модели заключается в объединении информации исходной модели с той, что получена в проверочном наборе данных,

и, как следствие, в улучшении результатов применения (*transportability*) такой модели для других людей.

Существует несколько методов обновления предсказательных моделей [2, 20, 31, 102, 290, 372, 373]. Методы различаются экстенсивно, что отражается в количестве повторно оцениваемых параметров. Как правило, наборы данных для разработки и проверки различаются частотой событий исхода, что приводит к плохой калибровке оригинальной модели на новых данных. Калибровку можно улучшить, если скорректировать свободный коэффициент (*intercept*) и исходные риски (если они известны) оригинальной модели для проверочной выборки, что потребует только один обновлённый параметр и, следовательно, небольшой набор проверочных данных [31, 290, 372, 373]. Более сложные методы обновления варьируют от общей корректировки всех весов предикторов с помощью одного коэффициента повторной калибровки, корректировки веса конкретного предиктора или добавления нового предиктора для переоценки всех коэффициентов регрессии. Последний метод применим, если набор проверочных данных значительно превышает количество данных, использованных для разработки (модели. — *Примеч. ред.*).

В **табл. 3** приведены различные методы обновления (моделей. — *Примеч. ред.*). Простые методы обновления (1 и 2) рассчитаны только на улучшение калибровки модели. Для улучшения дискриминации необходимы методы 3–6. Тем не менее обновлённые модели, особенно когда они получены на относительно небольших наборах проверочных данных, по-прежнему нуждаются в проверке перед применением в обычной практике [20].

И, наконец, как отмечено во **вставке В**, не рекомендуется обновлять существующую модель с использованием нового набора данных без предварительной количественной оценки предсказательной эффективности модели с новыми данными [47]. Если модель была обновлена, авторы должны обосновать необходимость обновления и описать, как это было сделано.

Группы риска

Пункт 11. Подробно опишите, как определяли группы риска (если применимо) (Р; П).

Примеры

«После того, как окончательная модель была определена, пациенты были разделены на группы риска двумя способами: 3 группы в соответствии с низким, средним и высоким риском (пороговые значения устанавливали в соответствии со значениями 25-го и 75-го процентилей распределения оценки риска, рассчитываемого моделью) и 10 групп с применением пороговых значений модели Кокса. Последнее минимизирует потерю информации для заданного количества групп. Поскольку в клинической практике использование 3 групп является общепринятым, для описания модели в дальнейшем использовали именно этот способ» [374]. (Прогнозирование; Разработка; Проверка.)

«Одна из целей этой модели — разработка доступного для клинициста метода стратификации риска для пациентов, готовящихся к операции по поводу злокачественного новообразования головы и шеи. С этой целью мы определили 3 категории риска переливания крови: низкий (<15%), промежуточный (15–24%) и высокий (>25%)» [375]. (Прогнозирование; Проверка.)

«Пациентов относили к группе высокого риска, если прогноз 10-летнего риска развития сердечно-сосудистых заболеваний составлял >20% в соответствии с рекомендациями NICE» [117]. (Прогнозирование; Проверка.)

«Выделили 3 группы риска на основе тертилей распределения ПИ (прогностический индекс. — *Примеч. авт.*). В подгруппе низкого риска (первый тертиль — ПИ <8,97) показатели бессобытийной выживаемости (БСВ; event-free survival) через 5 и 10 лет составили 100 и 89% (95% ДИ 60–97%) соответственно. В подгруппе промежуточного риска (второй тертиль — 8,97 <ПИ <10,06) показатели БСВ через 5 и 10 лет составляли 95% (95% ДИ 85–98%) и 83% (95% ДИ 64–93%), соответственно. В группе высокого риска (третий тертиль — ПИ >10,06) показатели БСВ через 5 и 10 лет составляли 85% (95% ДИ 72–92%) и 44% (95% ДИ 24–63%), соответственно» [376]. (Прогнозирование; Разработка.)

«В итоге вывели диагностическое правило с использованием уменьшенных округлённых коэффициентов многофакторной модели для оценки вероятности наличия сердечной недостаточности в диапазано от 0 до 100%. Пороговые значения для подтверждения и исключения сердечной недостаточности определили на основе клинически приемлемой вероятности ложноположительных (20 и 30%) и ложноотрицательных (10 и 20%) диагнозов» [377]. (Диагностика; Разработка; Проверка.)

Пояснение

Во многих исследованиях предсказательных моделей группы риска определяют на основании вероятностей, рассчитываемых многофакторной моделью. При представлении результатов или для облегчения принятия клинического решения чаще всего их обозначают как группы низкого, промежуточного (среднего) и высокого риска (пункты 3а и 20).

Нет единого мнения о том, как определять группы риска и сколько их должно быть [43]. Определение групп риска необходимо сопроводить описанием их границ (т.е. диапазона предсказываемых вероятностей для каждой группы) и способа их выбора. Однако если категоризация риска выполнена для помощи в принятии решений, авторам следует обосновать количество групп риска и выбор пороговых значений риска.

Есть опасения, что использование групп риска может не отвечать интересам пациентов [2, 112]. Такая категоризация риска хотя и является произвольной, может стать стандартом, несмотря на отсутствие какого-либо обоснования (например, в случае Ноттингемского прогностического индекса, *Nottingham Prognostic Index*) [378]). Кроме

того, упрощение предсказаний означает, что риски (вероятности) будут одинаковыми для всех пациентов в каждой категории. Поэтому независимо от определения каких-либо групп риска отчёты должны содержать достаточно информации (свободный коэффициент и коэффициенты бета логистической регрессионной модели, номограммы или веб-калькуляторы для детальных и более сложных вычислений), чтобы можно было рассчитать не только групповые риски (*group-based risks*), но и риски для конкретных лиц (*subject-specific risks*) (пункт 15a).

В некоторых случаях группы риска могут быть сформированы на основе внешних данных, которые предлагают другой план лечения или ведения, основанный на определённых пороговых значениях риска (например, показан ли статин для предотвращения сердечно-сосудистого события в случае, если прогностический риск выше или ниже определённого порогового значения [117]). Однако в большинстве случаев такие явные указания, основанные на предполагаемых вероятностях, отсутствуют.

Обзор 47 предсказательных моделей в онкологии показал, что группы риска были определены в 36 (76%) исследованиях, но подход к определению групп был неясен или не описан в 17 (47%) исследованиях [54]. В других обзорах авторы пришли к аналогичным выводам [43].

Разработка против проверки

Пункт 12. В проверочном исследовании укажите любые отличия в условиях проведения, критериях отбора, исходе и предикторах от таковых в исследовании, в котором модель была разработана (П).

Примеры

«...Суммарная величина риска по шкале GRACE соответствует предполагаемой вероятности общей смертности (*all-cause mortality*) в течение 6 месяцев после выписки из больницы. <...> Применимость оценки риска на период свыше 6 месяцев неизвестна. Цель исследования — изучить, будет ли показатель риска по шкале GRACE, рассчитанный при выписке из больницы, прогнозировать долгосрочную (до 4 лет) смертность в отдельной когорте регистра...» [379]. (Прогнозирование; Другой исход.)

«Правило Уэллса (*Wells rule*) разработано на данных, полученных от пациентов с подозрением на тромбоз глубоких вен, которые обратились в специализированные амбулаторные клиники. Хотя распространено мнение, что пациенты специализированных амбулаторных клиник схожи с таковыми в клиниках первичного звена здравоохранения, различия могут быть обусловлены механизмом направления этих пациентов врачами первичной помощи. Истинная диагностическая или дискриминационная точность правила Уэллса формально ранее не была подтверждена у пациентов с подозрением на ТГВ (тромбоз глубоких вен. — *Примеч. ред.*), наблюдавшихся в учреждениях первичной медицинской помощи. Необходимость проведения проверочного исследования продиктована тем, что эффективность любого диагностического или прогностического предсказывающего правила (*prediction rule*), скорее всего, будет ниже, чем ожидалось на основании данных исходного исследования, если оно применяется к новым пациентам, особенно когда эти пациенты выбираются в других условиях (наблюдения. — *Примеч. ред.*). Наша цель — количественно оценить диагностическую эффективность правила Уэллса в отношении пациентов, получающих первичную медицинскую помощь, и сравнить её с результатами, полученными в оригинальных исследованиях Уэллса и его коллег» [188]. (Диагностика; Разные условия.)

«В случаях, когда определения переменных в различных исследованиях не совпадали (например, физическая активность), мы использовали наилучшие доступные определения этих переменных для достижения разумной согласованности между базами данных. Например, в исследовании NHANES (National Health and Nutrition Examination Survey. — *Примеч. ред.*) мы классифицировали участников как физически активных, если в ответ на «Сравните свою активность с другими сверстниками» они указывали «более активен». В противном случае классифицировали участников как физически неактивных. В исследовании ARIC (Atherosclerosis Risk in Communities. — *Примеч. ред.*) физическая активность оценивалась в вопросе с ответом «да» или «нет», а в исследовании CHS (Cardiovascular Health Study. — *Примеч. ред.*) мы предложили два варианта ответов на вопрос о физической активности: «нет» или «низкая» против «умеренная» или «высокая»» [380]. (Прогнозирование; Разные предикторы.)

«Поскольку в исследовании NWAHS (The North West Adelaide Health Study. — *Примеч. ред.*) данные о применении антигипертензивных препаратов не собирали, мы предположили, что никто из участников не принимал таких препаратов. Аналогично в исследовании BMES (Blue Mountains Eye Study. — *Примеч. ред.*) не собирали данные о высоком уровне глюкозы в крови в анамнезе. На этом основании мы предположили, что ни у одного из участников таких случаев в анамнезе не было» [381]. (Прогностический; Разные предикторы.)

Пояснение

Описывая исследования эффективности предсказательной модели на различных наборах данных, авторы должны ясно и однозначно определить любые различия, запланированные или нет, которые могут потенциально повлиять на применение модели в других условиях [26, 28].

Предсказательные модели, разработанные в одних условиях оказания медицинской помощи (например, первичная медицинская помощь) или в конкретной стране, не обязательно одинаково полезны в других условиях (например, специализированная медицинская помощь) или в другой стране [19–21, 26, 28, 33, 183, 382, 383]. Например, случаи (пункт 5a) в условиях оказания специализированной медицинской помощи отличаются бóльшим количеством признаков и симптомов (и более узкими диапазонами значений предикторов), а также более тяжёлым

статусом заболевания в сравнении со случаями в учреждениях первичного звена здравоохранения [20, 21, 102].

Критерии отбора могут также различаться незапланированно (например, более широкий или ограниченный возрастной диапазон), что приводит к некоторым отличиям в наблюдениях [186], или даже запланировано (например, проверка предсказательной модели, которая была разработана для взрослых пациентов, в детской популяции [191, 384]).

В проверочном исследовании исход может быть таким же, как и в исследовании по разработке модели, но точное определение или метод измерения исхода может быть другим. Например, сахарный диабет можно определить по содержанию глюкозы в крови натощак, с помощью перорального глюкозотолерантного теста или исходя из сообщаемой пациентом информации о наличии у него заболевания [380, 385]. Даже если определение и методы измерения исхода одинаковы, отличия могут быть вызваны разными условиями проведения исследования, например, разной квалификацией наблюдателей (например, радиологов или патологоанатомов), разными лабораторными процедурами или технологиями визуализации.

Как и в случае с условиями и критериями отбора, отличия в исходах также могут быть запланированными. Целью исследования может быть оценка применимости модели для прогнозирования другого исхода [379, 383, 386]. Так, модели, разработанные для прогнозирования риска смерти после операции на сердце, были исследованы на предмет прогнозирования длительности пребывания в отделении интенсивной терапии [46]. Кроме того, существующие предсказательные модели могут быть оценены для предсказания наступления одинаковых исходов, но в разные моменты времени [387]. Например, модель GRACE, предсказывающую 6-месячную смертность у пациентов с острым коронарным синдромом [388], впоследствии использовали для прогнозирования смертности на протяжении 4 лет [379].

Наконец, могут отличаться определение и измерение предикторов, опять же намеренно или нет. Когда определения одинаковы, разница может объясняться изменением условий измерения предиктора. Например, специфический показатель крови может быть первоначально определён лабораторным методом в венозной крови, но проверен с точки зрения применимости с помощью экспресс-теста капиллярной крови [136, 389].

Авторы проверочных исследований также должны чётко указывать, как кодировали предикторы, т.е. приводить единицы измерения для всех непрерывных предикторов и критерии определения категориальных предикторов (например, для переменной «пол» женщины кодируются значением 0, мужчины — 1); см. пункты 7a и 15a. Более того, при использовании исторических данных для оценки эффективности предсказательной модели сведения о предикторе в этом наборе могут отсутствовать, поскольку данные собирали для другой цели. По этой причине исследователи могут использовать альтернативные предикторы (*proxy predictors*) [46], выполнить замещение отсутствующих данных или исключить предиктор из модели [198]. Последнего (эквивалентно присваиванию предиктору нулевого значения) следует избегать, поскольку предсказания модели в проверочном наборе данных будет сложно интерпретировать [198] (как в случае с моделью FRAX) [312, 314].

Поэтому важно, чтобы авторы проверочных исследований ясно и однозначно сообщали о том, имели ли место (запланированные или нет) изменения в условиях проведения, критериях отбора, предикторах, определении и измерении исхода, или включали в отчёт заявление о том, что условия, определения и измерения в их работе идентичны тем, которые были в исследованиях по разработке модели. Они должны не просто перечислить критерии отбора, исход и предикторы, но ясно и однозначно выделить любые различия и способы их устранения.

В 6 из 45 исследований (13%) по внешней проверке (*external validation*) модели, включённых в недавно опубликованный систематический обзор, было неясно, соответствует ли определение исхода его оригинальному определению [122].

Результаты

Участники

Пункт 13a. Опишите поток участников в ходе исследования, включая количество участников с исходом и без него, и, если применимо, характеристики периода отслеживания исходов. Графическое представление этой информации может быть полезным (P; П).

Примеры. Поток участников.

См. **рис. 3** и **4**.

Примеры. Период отслеживания исходов (*follow-up time*)

«Мы рассчитали 10-летний предполагаемый сердечно-сосудистый риск для каждого пациента в когорте THIN по шкале QRISK2-2011 … и отследили исходы у 292 928 (14,1%) пациентов в течение 10 лет и более» [117]. (Прогнозирование; Проверка.)

«На момент анализа 204 (66%) пациента умерли. Медиана продолжительности наблюдения за выжившими пациентами составила 12 месяцев (диапазон 1–84)» [391]. (Прогнозирование; Разработка.)

«Медиану продолжительности наблюдения рассчитывали в соответствии с обратным методом Каплана–Мейера (“*reverse Kaplan Meier*”), который рассчитывает потенциальный период наблюдения таким же образом, как и оценку функции выживаемости Каплана–Мейера, но с обратным значением индикатора состояния. Таким образом, смерть цензурирует истинное, но неизвестное время наблюдения за человеком, и цензурирование является конечной точкой (Schemper и Smith, 1996)» [392]. (Прогнозирование; Разработка.)

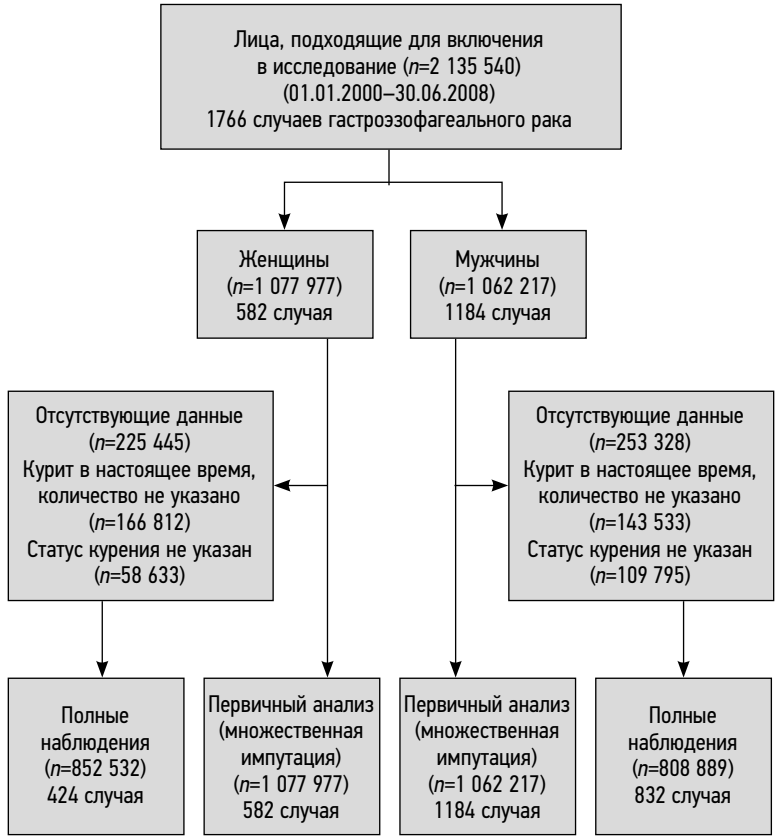


Рис. 3. Пример рисунка: схема изменения состава участников.

Примечание. Перепечатано из [390] с разрешения Elsevier.

Рис. 4. Пример рисунка: схема изменения состава участников.

Примечание. Воспроизведено из ссылки [377] с разрешения. NT-proBNP — N-концевой предшественник мозгового натрийуретического пептида.

Пояснение

Читателям важно сообщить об источнике выборки исследования, в том числе способе её формирования из большей первоначальной группы. Такая информация принципиальна для оценки контекста, в котором предсказательная модель может быть проверена или применена. Хотя описание потока участников в процессе исследования может быть представлено в тексте или в таблице, потоковые диаграммы (*flow diagram*) являются ценным способом, чтобы наглядно показать происхождение выборки исследования, с использованием данных которой модель была разработана или проверена.

Исходной точкой схематичного описания является указание на источник потенциальных участников, последующие шаги в схеме могут связывать критерии отбора и доступность данных [108] (пункт 5б). Дополнительно можно указать число участников с отсутствующими данными и количество событий исхода.

Для прогностических исследований важно описать продолжительность наблюдения для включённых участников, которое часто представляют с указанием медианы времени. Метод вычисления медианы продолжительности наблюдения должен быть отмечен. Один из таких способов — обратный метод Каплана–Мейера, который анализирует данные всех пациентов в когорте [393]. В примере выше использован стандартный метод Каплана–Мейера, но с обратным значением индикатора исхода, так что целевым исходом становится цензурированное наблюдение [108]. Также может быть полезно указать медиану продолжительности наблюдения за теми пациентами, у которых не наступило событие исхода (т.е. за пациентами с цензурированным временем выживания). Для моделей, предсказывающих вероятность события в конкретный момент времени, полезно сообщить о количестве лиц, за которыми наблюдали до этого момента времени.

Для диагностических исследований с отсроченной верификацией заболевания в качестве исхода (пункты 4а и 6б) также важно сообщать о медиане продолжительности наблюдения. Если данные исследования были разделены на наборы для разработки и проверки модели, полезно предоставить всю вышеуказанную информацию для каждой выборки.

Недавние систематические обзоры исследований предсказательных моделей показали, что многие авторы не указывают количество событий исхода [34, 45, 54, 85, 394]. В других обзорах отмечено, что в исследованиях часто не приводится описание продолжительности последующего наблюдения [43].

Пункт 13б. Опишите характеристики участников исследования (основные демографические и клинические показатели, доступные предикторы), укажите количество участников с отсутствующими данными по показателям предикторов и исхода (Р; П).

Примеры

См. **табл. 5** и **6**.

Пояснение

Чёткое описание распределения [распространённость (*prevalence*), среднее или медиана, стандартное отклонение или межквартильный размах] важных характеристик участников исследования необходимо для оценки контекста, выборки и условий проведения исследования. На основании этой информации читатели смогут судить о том, можно ли в принципе проверить предсказательную модель на своих данных или применить к своим пациентам. Недостаточно сообщить только критерии включения в исследование. В идеале необходимо сообщить обо всех предикторах, особенно тех, которые включены в окончательную модель, а также других важных переменных (демографические, клинические данные, сведения об условиях наблюдения). Кроме того, следует указывать диапазоны значений всех количественных предикторов, особенно в окончательной модели. В отсутствии такой информации неясно, к кому может быть применима модель (пункт 15а).

Вышеуказанную информацию наиболее эффективно отображать в таблице, которая также должна включать количество (процент) отсутствующих наблюдений для каждой переменной (см. **табл. 4**). Если наблюдения отсутствуют только для нескольких переменных, об этом можно сообщить в тексте отчёта.

Полезно также включить описательную информацию об исходе и, если проводится однофакторный анализ, показать сводную статистику предикторов и других важных для исследования переменных по различным категориям исходов (пункт 14б). В качестве альтернативы можно показать частоту исходов для категорий предикторов.

Нет никаких доказательств того, что отчётность о характеристиках участников или предикторах является особенно плохой. Однако в нескольких систематических обзорах выявлены исследования, в которых такая ключевая информация не была представлена [43, 62, 71, 72, 122]. В недавнем обзоре 78 исследований, посвящённых внешней проверке эффективности 120 предсказательных моделей, диапазон непрерывных предикторов был указан лишь в 8% (10 из 120) оригинальных исследований, в которых разрабатывалась оцениваемая предсказательная модель [122].

Пункт 13в. Для проверочных исследований – представьте сравнение распределения важных переменных (демографические показатели, предикторы, исход) с данными, использованными для разработки модели (П).

Примеры

См. **табл. 7** и **8**.

Пояснение

В проверочное исследование предсказательной модели обычно включают участников, схожих с теми, которые были использованы в оригинальном исследовании

Таблица 5. Пример таблицы. Характеристики участников исследования

Характеристика	Отсутствующие данные, *n* (%)	Значение
Пациенты с подтверждённой ТЭЛА	0	222 (23,0%)
Общие характеристики		
Средний возраст	0	60,6 лет (СО 19,4)
Средняя масса тела	83 (8,6)	72,6 кг (СО 16,1)
Мужчины	0	403 (41,8%)
Факторы риска		
Пациенты с семейным анамнезом ТГВ или ТЭЛА	6 (0,6)	102 (10,6%)
Пациенты с ТГВ или ТЭЛА в анамнезе	2 (0,2)	166 (17,2%)
Пациенты с подтверждённой хронической сердечной недостаточностью	0	95 (9,8%)
Пациенты, перенёсшие инсульт	0	29 (3,0%)
Пациенты с ХОБЛ	0	99 (10,3%)
Пациенты, перенёсшие хирургическое вмешательство и (или) перелом в последний месяц	0	67 (6,9%)
Пациенты, обездвиженные в течение последнего месяца	0	165 (17,1%)
Пациенты с прогрессирующим онкологическим заболеванием	3 (0,3)	89 (9,2%)
Пациенты, принимающие оральные контрацептивы в настоящее время	1 (0,1)	69 (7,2%)
Пациентки с беременностью или после родов	0	10 (1,0%)
Симптомы		
Пациенты с обмороками	2 (0,2)	68 (7,0%)
Пациенты с недавним кашлем	0	197 (20,4%)
Пациенты с кровохарканьем	0	43 (4,5%)
Пациенты с одышкой	0	637 (66,0%)
Пациенты с болью в грудной клетке	0	681 (70,6%)
Пациенты с односторонней болью в нижней конечности	0	138 (14,3%)
Клиническое обследование		
Общие признаки:		
Средняя температура тела	37 (3,8)	36,9 °C (СО 0,8)
Средняя частота сердечных сокращений	4 (0,4)	86,3 уд/мин (СО 19,7)
Средняя частота дыхания	59 (6,1)	20,2 цикла/мин (СО 7,0)
Среднее систолическое артериальное давление	6 (0,6)	140 мм рт. ст. (СО 23)
Среднее диастолическое артериальное давление	7 (0,7)	81 мм рт. ст. (СО 15)
Признаки ТЭЛА:		
Пациенты с хронической венозной недостаточностью	3 (0,3)	199 (20,6%)
Пациенты с варикозным расширением вен	15 (1,6)	227 (23,5%)
Пациенты с односторонним отёком и болью при пальпации глубоких вен	0	51 (5,3%)
Пациенты с признаками патологии при аускультации грудной клетки	2 (0,2)	158 (16,4%)
Пациенты со вздутием шейных вен	2 (0,2)	108 (11,2%)

Примечание. СО — стандартное отклонение, ХОБЛ — хроническая обструктивная болезнь лёгких; ТГВ — тромбоз глубоких вен; ТЭЛА — тромбоэмболия лёгочной артерии. Заимствовано из источника [395].

Таблица 6. Пример таблицы. Характеристики участников исследования

Характеристика	Все пациенты (*n*=202)	ТБ есть* (*n*=72)	ТБ нет (*n*=130)	*P*
Медиана возраста (МКР), *годы*	32 (28–39)	32 (28–39)	33 (28–40)	0,59
Женский пол, *%*	113 (56)	38 (53)	75 (58)	0,50
Впервые диагностированная ВИЧ-инфекция, %	53 (26)	14 (19)	39 (30)	0,10
Медиана количества CD4 (МКР), *клеток/мкл†*	64 (23–191)	60 (70–148)	74 (26–213)	0,17
Принимают ко-тримоксазол в профилактических целях, %‡	117 (58)	48 (67)	69 (53)	0,061
Принимают антиретровирусную терапию, %§	36 (18)	15 (21)	21 (16)	0,41
Принимали антибиотики до госпитализации, %	134 (66)	51 (71)	83 (64)	0,31
Летальность двух-месячная, *%*¶	58 (32)	27 (42)	31 (26)	0,028

Примечание. МКР — межквартильный размах; ТБ — туберкулез. Заимствовано из источника [396]. * Подтвержден положительными результатами посева (обнаружение культуры микобактерий) мокроты или бронхоальвеолярного лаважа на твердую питательную среду. † В 4 случаях результаты отсутствуют. ‡ Все, кроме 1 пациента, принимали ко-тримоксазол ⩾ 1 месяца. § Все пациенты сообщили о приёме антиретровирусных препаратов в течение ⩾1 месяца. 8 пациентов с ТБ и 12 пациентов без ТБ выбыли из-под наблюдения.

Таблица 7. Пример таблицы. Сравнение характеристик участников, данные которых использованы при разработке и проверке модели (Разработка; Проверка)

Характеристика	Когорта для создания модели (*n*=8820)	Когорта для внутренней проверки (*n*=5882)	Когорта для внешней проверки (*n*=2938)
Демографические показатели			
Медиана возраста (МКР), годы	66 (56–74)	66 (57–75)	64 (55–72)
Мужской пол, n (*%*)	5430 (61,6)	3675 (62,5)	1927 (65,5)
Сосудистые факторы риска, n (*%*)			
Гипертензия	5601 (63,5)	3683 (62,6)	1987 (67,6)
Сахарный диабет	1834 (20,8)	1287 (21,9)	720 (24,5)
Дисплипидемия	947 (10,7)	637 (10,8)	386 (13,1)
Фибрилляция предсердий	643 (7,3)	415 (7,1)	175 (6,0)
Ишемическая болезнь сердца	1222 (13,9)	811 (13,8)	285 (9,7)
Заболевания периферических артерий	64 (0,7)	29 (0,5)	26 (0,9)
Инсульт/ТИА в анамнезе	2795 (31,7)	1822 (31,0)	809 (27,5)
Курение	3510 (39,8)	2326 (39,5)	1022 (34,8)
Чрезмерное потребление алкоголя	1346 (15,3)	921 (15,7)	372 (12,7)
Другие сопутствующие заболевания, *n (%)*			
Хроническая сердечная недостаточность	169 (1,9)	121 (2,1)	24 (0,8)
Порок сердца	213 (2,4)	139 (2,4)	40 (1,4)
Хроническая обструктивная болезнь лёгких	98 (1,1)	64 (1,1)	12 (0,4)
Цирроз печени	29 (0,3)	21 (0,4)	7 (0,2)
Язвенная болезнь (желудка, двенадцатиперстной кишки) или ЖКК	283 (3,2)	195 (3,3)	76 (2,6)
Почечная недостаточность	7 (0,1)	4 (0,1)	3 (0,1)
Артрит	266 (3,0)	176 (3,0)	45 (1,5)
Деменция	113 (1,3)	82 (1,4)	18 (0,6)
Онкологические заболевания	150 (1,7)	109 (1,9)	54 (1,8)
Ограничение жизнедеятельности до инсульта (МШР ⩾3), *n (%)*	809 (9,2)	535 (9,1)	0 (0,0)
Антитромбоцитарная терапия до госпитализации, *n (%)*	1449 (16,4)	932 (15,8)	357 (12,2)
Антикоагулянтная терапия до госпитализации, *n (%)*	210 (2,4)	122 (2,1)	26 (0,9)
Медиана оценки по шкале NIHSS при госпитализации (МКР)	5 (2–9)	5 (2–9)	4 (2–8)
Медиана оценки по шкале GCS при госпитализации (МКР)	15 (14–15)	15 (14–15)	15 (15–15)
Медиана САД при госпитализации (МКР), *мм рт. ст.*	150 (134–163)	150 (135–162)	150 (135–167)

Медиана ДАД при госпитализации (МКР), *мм рт. ст.*	89 (80–95)	89 (80–95)	90 (80–98)
Подтип инсульта по классификации OCSP, *n (%)*			
Парциальный инфаркт в бассейне внутренней сонной артерии	4834 (54,8)	3327 (56,6)	1829 (62,3)
Обширный инфаркт в бассейне внутренней сонной артерии	811 (9,2)	519 (8,8)	176 (6,0)
Лакунарный инфаркт	1667 (18,9)	1074 (18,3)	246 (8,4)
Инфаркт в вертебрально-базилярном бассейне	1508 (17,1)	962 (18,4)	687 (23,4)
tPA внутривенно в течение 3 ч после клинического события, *n (%)*	108 (1,2)	73 (1,2)	137 (4,6)
Антитромботическая терапия при госпитализации, *n (%)*	7371 (83,6)	4950 (84,2)	2550 (86,8)
Антикоагулянтная терапия при госпитализации, *n (%)*	210 (2,4)	122 (2,1)	159 (5,4)
Медиана продолжительности пребывания в стационаре (МКР), *сутки*	14 (10–20)	14 (10–20)	14 (11–18)
ЖКК в стационаре, *n (%)*	227 (2,6)	135 (2,3)	44 (1,5)

Примечание. ДАД — диастолическое артериальное давление; GCS — шкала комы Глазго; ЖКК — желудочно-кишечное кровотечение; МКР — межквартильный размах; МШР — модифицированная шкала Rankin; NIHSS — шкала инсульта Национальных институтов здоровья; OCSP — Проект по борьбе с инсультом в Оксфордшире; САД — систолическое артериальное давление; ТИА — транзиторная ишемическая атака; tPA — тканевой активатор плазминогена. Заимствовано из источника [397].

Таблица 8. Пример таблицы. Сравнение характеристик участников, данные которых использовали для разработки и проверки модели (Проверка)

Предиктор риска	QRESEARCH		THIN (Внешняя проверка)*		
	Разработка (*n*=2 355 719)	Внутренняя проверка (*n*=1 238 971)	Женщины (*n*=1 077 977)	Мужчины (*n*=1 062 217)	Всего (*n*=2 140 194)
Медиана возраста (СО), *годы*	50,1 (15,0)	50,1 (15,0)	49 (15,1)	47 (14,2)	48 (14,7)
Статус курения, *n (%)*					
Некурящие	1 194 692 (50,7)	624 788 (50,4)	477 785 (44,3)	369 315 (34,8)	847 100 (39,6)
Курильщик в прошлом	427 246 (18,1)	229 516 (18,5)	123 037 (11,4)	155 961 (14,7)	278 998 (13,0)
Курильщик в настоящее время, количество выкуриваемых сигарет не указано	71 416 (3,0)	39 231 (3,2)	166 812 (15,5)	143 533 (13,5)	310 345 (14,5)
Курение минимальное (<10 сигарет/сутки)	148 063 (6,3)	79 844 (6,4)	70 298 (6,5)	66 858 (6,3)	137 156 (6,4)
Курение умеренное (10–19 сигарет/сутки)	179 931 (7,6)	95 754 (7,7)	106 203 (9,9)	102 868 (9,7)	209 071 (9,8)
Курение интенсивное (≥20 сигарет/сутки)	133 980 (5,7)	73 554 (5,9)	75 209 (7,0)	113 887 (10,7)	189 096 (8,8)
Нет данных	200 391 (8,5)	96 284 (7,8)	58 633 (5,4)	109 795 (10,3)	168 428 (7,9)
Симптомы в настоящее время и за предыдущий год, *n (%)*					
Дисфагия в настоящее время	15 021 (0,6)	8165 (0,7)	10 391 (1,0)	8846 (0,8)	19 237 (0,9)
Гематемезис в настоящее время	12 952 (0,5)	7119 (0,6)	4630 (0,4)	6162 (0,6)	10 792 (0,5)
Боль в животе в настоящее время	225 543 (9,6)	126 161 (10,2)	144 266 (13,4)	102 732 (9,7)	246 998 (11,5)
Потеря аппетита в настоящее время	9978 (0,4)	6133 (0,5)	3317 (0,3)	2521 (0,2)	5838 (0,3)
Потеря веса в настоящее время	9998 (0,4)	5377 (0,4)	15 465 (1,4)	12 938 (1,2)	28 403 (1,3)
Гемоглобин <11 г/дл в течение последнего года	22 576 (1,0)	12 638 (1,0)	13 792 (1,3)	4563 (0,4)	18 355 (0,9)

Примечание. THIN — The Health Improvement Network. *В когорте THIN большее число пациентов сообщали о боли в животе и потере массы тела, по сравнению с исходной когортой, данные которой использовали для разработки модели. Заимствовано из источника [390].

при разработке модели [19, 20, 26, 28, 33]. Однако, как уже обсуждалось в пункте 12, отличие популяции проверочного исследования от таковой в исследовании, в котором модель была разработана, может быть запланировано. В этой связи важно представить демографические характеристики, предикторы модели и исходы участников (проверочного) исследования наряду с теми, о которых сообщалось в оригинальной работе. Наиболее эффективно такая информация может быть представлена в таблице с демонстрацией распределения этих переменных в общей выборке и, если необходимо, в особых группах участников (например, сформированных с учётом пола). Также полезно указать количество отсутствующих наблюдений для каждой из упомянутых переменных

в обоих наборах данных (в оригинальном и проверочном исследованиях. — *Примеч. ред.*).

Можно возразить, что для хорошо известных и давно существующих моделей (например, шкал риска APACHE или Framingham) такое сравнение будет излишним. Однако не все читатели могут быть хорошо знакомы с этими моделями, поэтому мы всё же рекомендуем провести сравнение между наборами данных, использованных для проверки и разработки, или, если возможно, даже с данными предыдущих проверочных исследований.

Наконец, авторы должны объяснить причины любых заметных различий между выборками проверочного и предыдущего исследований, если это не было запланировано (пункт 12), и затем обсудить в статье возможные последствия этих расхождений для полученных результатов, таких как предсказательная эффективность модели в проверочном наборе данных (пункты 16 и 18).

Недавний систематический обзор 78 исследований с проверкой модели на независимых данных (включая исследования, в которых разработка модели была дополнена внешней проверкой), показал, что только в 31 (40%) отчёте сравнивали или обсуждали характеристики когорт оригинального исследования, в котором была разработана модель, и проверочного исследования с независимыми данными [122].

Разработка модели

Пункт 14a. Укажите количество участников и событий исхода для каждого анализа (P).

Примеры

См. **табл. 9** и **10**.

Пояснение

Как отмечено в пункте 8, эффективный размер выборки в исследованиях вопросов предсказания определяется количеством событий, а не участников. Отношение количества участников с событием к количеству исследуемых предикторов играет ведущую роль в оценке риска переобучения (*overfitting*) в конкретном исследовании (пункты 8 и 10б).

При отсутствующих данных количество участников и событий часто будет варьировать от анализа к анализу, если только участники с отсутствующими данными не будут исключены или такие данные не будут восстановлены путём подстановки (*imputation*) (пункт 9). При создании новой предсказательной модели авторы часто проводят анализ для изучения нескорректированной ассоциации (обычно называемой однофакторной или двухфакторной ассоциацией, *univariable* или *bivariable association*) между предиктором и исходом (пункт 14б). В этих случаях, если у участников отсутствуют какие-либо данные и их по этой причине исключат из анализа (парное удаление),

Таблица 9. Пример таблицы. Размер выборки и количество событий исхода (сравнение моделей)*

	Модель A		Модель B	
	Мужчины	**Женщины**	**Мужчины**	**Женщины**
Оценки когорты для разработки модели				
N	13 240	15 311	12 075	13 935
Количество событий	466	215	425	189
	Бета	**Бета**	**Бета**	**Бета**
Возраст (1 год)	0,053	0,080	0,241	0,066
Курение	0,466	0,776	2,453	0,784
Индекс массы тела	-	-	-	-
Диабет	-	-	0,528	0,778
САД (10 мм рт. ст.)	-	-	0,888	0,038
Общий холестерин (10 мг/дл)	-	-	0,061	0,077
Холестерин ЛПВП (10 мг/дл)	-	-	–0,211	–0,272
Лечение гипертензии при САД >120 мм рт. ст.	-	-	0,519	0,133
Продолжительность курения	-	-	–0,034	-
Продолжительность высокого САД	-	-	–0,013	-
10-летняя бессобытийная выживаемость по Коксу (%)	96,2	98,7	96,9	99,0
C-индекс	66,3	72,0	72,0	76,7
Оценка модели по данным проверочной когорты				
N	7955	9481	7955	9481
Количество событий	263	147	263	147
C-индекс	66,0	69,6	71,0	73,8

Примечание. ЛПВП — липопротеины высокой плотности; САД — систолическое артериальное давление. Заимствовано из источника [398]. * β-коэффициенты для переменных, включённых в упрощённую (A) и полную (B) модели, получены на данных исходной когорты для инфаркта миокарда или стенокардии, оценка эффективности модели выполнена на данных проверочной когорты с разбивкой по полу.

Таблица 10. Пример таблицы. Количество событий в каждом нескорректированном анализе

Характеристики	Пациенты с инфекцией *CD (n=395), n (%)*	Тяжёлое течение, вызванное инфекцией *CD, n (%)**		Отношение шансов (95% ДИ)	*P*
		Да	Нет		
Демографические показатели					
Возраст					
≤49 лет	85 (22)	6 (13)	79 (23)	1 (референсная категория)	0,01
50–84 года	275 (70)	31 (67)	237 (70)	1,72 (0,69–4,28)	
≥85 лет	35 (9)	9 (20)	23 (7)	5,15 (1,66–16,0)	
Мужской пол	220 (56)	24 (52)	191 (56)	0,85 (0,46–1,57)	0,59
Университетская клиника	266 (67)	23 (50)	239 (71)	0,42 (0,22–0,28)	0,01
Отделение диагностики					
Другие отделения	293 (74)	35 (76)	251 (74)	1 (референсная категория)	<0,01
Хирургическое отделение	83 (21)	4 (9)	78 (23)	0,37 (0,13–1,07)	
Отделение интенсивной терапии	19 (5)	7 (15)	10 (3)	5,02 (1,80–14,0)	
История терапевтических и хирургических вмешательств†					
Цитостатики	64 (16)	7 (15)	55 (16)	0,91 (0,39–2,15)	0,84
Иммунодепрессанты	172 (44)	21 (47)	146 (44)	1,13 (0,60–2,10)	0,71
Ингибиторы протонного насоса	251 (64)	34 (76)	211 (63)	1,82 (0,89–3,71)	0,10
Недавняя операция на брюшной полости	110 (28)	4 (9)	105 (31)	0,21 (0,07–0,59)	<0,01
Недавняя госпитализация	210 (55)	28 (61)	177 (54)	1,37 (0,71–2,49)	0,38
Антибиотики	335 (85)	34 (74)	293 (87)	0,44 (0,21–0,90)	0,03
Клинические показатели					
Индекс Charlson					
0	59 (15)	7 (15)	52 (15)	1 (референсная категория)	0,53
1–2	150 (38)	14 (30)	134 (40)	0,78 (0,30–2,03)	
3–4	120 (31)	15 (33)	101 (30)	1,10 (0,42–2,87)	
≥5	64 (16)	10 (22)	50 (15)	1,49 (0,53–4,21)	
Диарея как причина госпитализации	104 (27)	23 (50)	78 (23)	3,31 (1,76–6,22)	<0,01
Диарея во время госпитализации	283 (72)	28 (61)	248 (74)	0,55 (0,29–1,04)	0,06
Лихорадка	208 (60)	25 (66)	174 (59)	1,36 (0,67–2,76)	0,40
Гипотензия	117 (30)	25 (63)	88 (30)	3,86 (1,94–7,68)	<0,01
Кровавый понос (макроскопический)	52 (15)	7 (16)	44 (15)	1,14 (0,48–2,71)	0,77
Лабораторные данные					
Содержание креатинина до начала диареи					
<90	199 (58)	17 (43)	178 (61)	1 (референсная категория)	0,05
≥90	109 (32)	16 (40)	89 (30)	1,88 (0,91–3,90)	
Диализ	33 (10)	7 (18)	25 (9)	2,93 (1,11–7,77)	

Примечание. CD — *Clostridium difficile.* Заимствовано из источника [399]. * Данные об исходе отсутствуют у 10 (2,5%) пациентов, поэтому максимальное количество пациентов с тяжёлым течением составляет 46, без тяжёлого течения — 339 человек. † Данные о принимаемых лекарственных препаратах и хирургических вмешательствах собраны за 3 месяца до начала диареи.

количество участников при анализе нескорректированной ассоциации (*unadjusted association*) между каждым предиктором и исходом будет варьировать. Следовательно, если сообщается об однофакторных ассоциациях, следует указывать количество участников с полными данными по каждому предиктору и соответствующее им количество событий.

Точно так же авторы могут создать или сравнить эффективность нескольких многофакторных моделей на одном и том же наборе данных. Например, одна модель может быть построена на широко доступных предикторах, а другая включает дополнительные предикторы, которые ограниченно доступны (например, результаты анализа крови). В этой связи важно знать размер выборки и количество исходов, использованных для создания всех моделей.

Читатели должны ясно и однозначно понимать, какие участники были включены в каждый анализ. В частности, для исследований, в которых разрабатывается новая предсказательная модель, информация о количестве событий, используемых для создания модели, позволяет рассчитать показатели переобучения, такие как EPV (*events per variable,* пункты 8 и 10б). Для исследований по разработке модели, в которых данные были разделены на два набора (для создания и проверки модели), важно сообщать количество участников и событий исхода для каждого из них.

Пункт 14б. Если применимо, укажите нескорректированные оценки ассоциации каждого потенциального предиктора и исхода (P).

Примеры

См. **табл. 11**.

Пояснение

Однофакторный анализ желателен для того, чтобы позволить читателю подтвердить ожидаемые предсказываемые связи, основанные на результатах предыдущих исследований, а также для наблюдения за различиями в предсказательной точности предиктора по данным нескорректированного (однофакторного) и скорректированного (многофакторного) анализа. По такому же принципу строится отчётность в этиологических (причинных) и нерандомизированных интервенционных исследованиях, авторы которых часто описывают так называемые грубые (*crude*) и скорректированные ассоциации (*adjusted associations*) [97, 401]. Нескорректированные результаты являются исходными, с которыми проводится сравнение скорректированных результатов окончательной многофакторной предсказательной модели.

В случае однофакторного анализа бинарных конечных точек (например, 30-суточной летальности) авторы должны указать отношения рисков (*risk ratios*) или шансов (*odds ratios*) вместе с доверительными интервалами. Аналогичным образом, если предсказываются исходы

Таблица 11. Пример таблицы. Нескорректированная ассоциация между предикторами и исходом*

Характеристики	Пациенты с исходом (*n*=399)	Пациенты без исхода (*n*=15 881)	Одномерное отношение шансов (95% ДИ)	Многомерное отношение шансов (95% ДИ)	*P*
Демографические данные					
Средний возраст (СО), *годы*	81 (8)	75 (8)	1,8 (1,6–1,9)	1,6 (1,4–1,8)	<0,001
Мужчины	41	38	1,2 (1,0–1,4)	1,3 (1,1–1,7)	0,008
Медицинская помощь в анамнезе					
Предыдущая госпитализация в связи с пневмонией или гриппом	16	1	22,4 (16,3–30,6)	8,1 (5,7–11,5)	<0,001
Среднее число амбулаторных посещений (СО), *n*	26 (27)	11 (14)	2,4 (2,1–2,7)	1,5 (1,3–1,8)	<0,001
Сопутствующие состояния					
Заболевание сердца	50	24	3,2 (2,6–3,8)	1,2 (1,0–1,5)	0,10
Заболевание лёгких	40	14	4,1 (3,3–5,0)	1,8 (1,4–2,3)	<0,001
Деменция или инсульт	31	9	4,6 (3,7–5,8)	2,1 (1,6–2,7)	<0,001
Заболевание почек	13	4	4,0 (2,9–5,4)	1,5 (1,1–2,1)	0,02
Онкологическое заболевание	12	2	6,8 (4,9–9,4)	4,9 (3,4–7,0)	<0,001
Диабет	19	12	1,8 (1,4–2,3)	-	-
Анемия	24	8	3,7 (2,9–4,7)	-	-
Дефицит питательных веществ	5	2	3,7 (2,4–5,9)	-	-
Васкулит или ревматическое заболевание	3	2	1,3 (0,7–1,3)	-	-
Иммунодефицит	2	1	2,0 (1,0–4,0)	-	-
Цирроз печени	1	0.3	3,1 (1,1–8,7)	-	-

Примечание. * Данные представлены в процентах, если не указано иное. Заимствовано из источника [400].

во времени (*time-to-event outcomes*), авторы также должны представить отношения рисков (*hazard ratios*) и соответствующие доверительные интервалы. Значения *P* могут быть представлены, хотя они не предоставляют дополнительной информации при наличии доверительных интервалов. Обычно такие результаты представляют в табличной форме, часто в сочетании с результатами (оценка ассоциации в парах предиктор — исход) многофакторного анализа.

При отсутствии данных авторы должны указать количество участников, включённых в каждый нескорректированный анализ (пункт 14а). Для дихотомических или категориальных предикторов авторы должны сообщить для каждой категории количество участников, у которых наступил изучаемый исход.

Однако вслед за другими авторами мы не рекомендуем включать в многофакторную модель предикторы исключительно на основании результатов их нескорректированной ассоциации с исходом [2, 112, 235] (пункт 10б).

Характеристики модели

Пункт 15а. Представьте полную предсказательную модель, позволяющую предсказывать исход для отдельных лиц (т.е. все коэффициенты регрессии и свободный коэффициент модели или исходный показатель выживаемости в определённый момент времени) (P)

Примеры

См. **табл. 12–14**.

Пояснение

Предсказательные модели должны быть описаны достаточно подробно, чтобы можно было делать предсказания для отдельных лиц либо для последующих проверочных исследований, либо для клинической практики (пункт 15б). В случае бинарных исходов необходимо

Таблица 12. Пример таблицы. Полная прогностическая модель (выживаемости), включая данные об исходной функции выживания в определённый момент времени*

	β-коэффициент	SE	Значение *P*
Возраст	0,15052	0,05767	0,009
Возраст^2	–0,00038	0,00041	0,35
Мужской пол	1,99406	0,39326	0,0001
Индекс массы тела	0,01930	0,01111	0,08
Систолическое артериальное давление	0,00615	0,00225	0,006
Лечение гипертензии	0,42410	0,10104	0,0001
PR-интервал	0,00707	0,00170	0,0001
Значимые шумы в сердце	3,79586	1,33532	0,005
Сердечная недостаточность	9,42833	2,26981	0,0001
Мужской пол × возраст^2	–0,00028	0,00008	0,0004
Возраст × значимые шумы в сердце	–0,04238	0,01904	0,03
Возраст × сердечная недостаточность	–0,12307	0,03345	0,0002

Примечание. * $S_0(10)$=0,96337 (10-летняя исходная выживаемость). Значения β соответствуют каждому увеличению значения непрерывных переменных на единицу измерения и каждому состоянию дихотомических переменных. Заимствовано из источника [402].

Таблица 13. Пример таблицы. Полная диагностическая (логистическая) модель, включая свободный коэффициент*

Свободный коэффициент и предикторы	β†	Отношение шансов	95% ДИ
Свободный коэффициент	–3,66		
Потомственный пекарь	0,67	2,2	1,2–3,9
Назальные и конъюнктивальные симптомы в последние 12 месяцев	0,72	2,3	1,2–4,5
Симптомы астмы в последние 12 месяцев	0,63	2,0	0,9–4,4
Одышка и хрипы	0,61	2,3	1,3–3,8
Симптомы верхних дыхательных путей, связанные с работой	0,47	1,7	0,9–3,1
Симптомы нижних дыхательных путей, связанные с работой	0,61	2,2	1,1–4,4
Площадь под ROC (95% ДИ)	0,75 (0,71–0,81)		

Примечание. ROC — receiver-operating characteristic. Заимствовано из источника [319]. * Предсказанную вероятность сенсибилизации к пшенице можно рассчитать по следующей формуле: *P(сенсибилизация)*=1/(1+ exp(–(–3,66 + потомственный пекарь × 0,67 + назальные/конъюнктивальные симптомы в последние 12 месяцев × 0,72 + симптомы астмы в последние 12 месяцев x 0,63 + одышка и хрипы × 0,61 + симптомы верхних дыхательных путей, связанные с работой × 0,47 + симптомы нижних дыхательных путей, связанные с работой x 0,61))). Значение предиктора равно 1, если соответствующее состояние присутствует, и 0, если оно отсутствует. † Коэффициент регрессии, умноженный на коэффициент сжатия (вычислен путём процедуры бутстреппинга), равный 0,89.

Таблица 14. Пример таблицы. Оригинальная и обновлённая предсказательные модели

Предиктор	Оригинальная модель	Обновлённая модель
Возраст (годы)	–0,022	–0,017
Женский пол	0,46	0,36
Курит в настоящее время	–0,63	–0,50
ПОТР или расстройства движения в анамнезе	0,76	0,60
Хирургия нижних отделов брюшной полости или среднего уха	0,61	–
Хирургия брюшной полости или среднего уха*	–	0,48
Анестезия изофлураном и (или) закисью азота†	0,72	–
Ингаляционная анестезия‡	–	0,35
Амбулаторная хирургия§	–	–1,16
Свободный коэффициент	0,15	0,12

Примечание. ПОТР — послеоперационная тошнота и рвота. Заимствовано из источника [187]. *В обновлённой модели… этот предиктор заменил показатель «хирургия нижних отделов брюшной полости или среднего уха» в оригинальной модели. Полное определение предиктора в обновлённой модели — «хирургия нижних и верхних отделов брюшной полости, лапароскопическая хирургия и хирургия среднего уха». † По сравнению с внутривенной анестезией пропофолом. ‡ По сравнению с внутривенной анестезией пропофолом. В обновлённой модели… этим определением заменили предиктор оригинальной модели «анестезия изофлураном и (или) закисью азота». § Предиктор, не включённый в оригинальную модель.

указывать коэффициент регрессии или отношение шансов (*odds ratio*) для каждого предиктора модели и свободный коэффициент (*intercept*). Хорошая общепринятая практика — указывать доверительные интервалы для каждого рассчитанного коэффициента [403], хотя в проверочных исследованиях или в клинической практике они не используются. Это относится и к параметрической модели выживаемости, прогнозирующей (для длительного периода) наступление исхода во времени. Если авторы применяли методы сжатия (*shrinkage methods*) (пункт 10б), следует указать исходные и уменьшенные коэффициенты регрессии.

Иных рекомендаций следует придерживаться в случае полупараметрической регрессионной модели Кокса (*semi-parametric Cox regression model*), часто используемой для предсказания исхода во времени. В этом случае авторы должны представить коэффициент регрессии или отношение рисков (*hazard ratio*) для каждого предиктора модели вместе с его доверительным интервалом. Однако модель Кокса не имеет свободного коэффициента, и индивидуальные вероятности выживания оцениваются относительно неопределённой исходной функции выживания (*baseline survival function*). Следовательно, вероятности не могут быть оценены только на основе коэффициентов регрессии.

Для оценки вероятности исхода у отдельных лиц в конкретный момент времени, авторы должны указать совокупный исходный риск (или исходную выживаемость) для одной или более клинически значимых временных точек (пункт 15б). В исследованиях сердечно-сосудистых или онкологических заболеваний часто выбирают 5- или 10-летнюю выживаемость, но возможны и другие временные точки. В качестве альтернативы авторы, разрабатывающие предсказательные модели с использованием метода регрессии Кокса, должны рассмотреть возможность оценки и представления исходной функции риска с использованием дробных полиномов (*fractional polynomials*) или ограниченных кубических сплайнов (*restricted cubic splines*) [297, 309, 373, 404].

Предоставление полной информации о сложной модели (например, модели ICNARC [405]) может оказаться непростой задачей. В других случаях модели регулярно обновляются и постоянно размещаются в Интернете, но не в журнальных статьях (например, QRISK2 [139]). Независимо от сложности или частоты обновления модели, мы настоятельно рекомендуем представлять полную модель в рецензируемой статье или в веб-приложении. Если детали модели остаются неопубликованными, она никогда не будет проверена, и в связи с этим весьма сомнительно, следует ли рассматривать такую модель для клинического использования [312, 314, 406].

Помимо сообщения точной формулы разработанной модели, необходимо указать, каким образом кодировали все предикторы (см. также пункт 7а). Для всех непрерывных предикторов следует привести шкалу измерений (например, измеряется ли окружность талии в сантиметрах или дюймах). Если непрерывные переменные разделены на категории (пункт 10а и **вставка Д**), следует указать пороговые значения для всех категорий, включая нижний и верхний предел первой и последней категории соответственно, о которых часто не сообщают. Для категориальных предикторов авторы должны чётко указать, как они были закодированы — например, при регистрации пола участника женщин кодировали как 0, а мужчин — как 1.

Кроме того, следует чётко указывать диапазоны всех непрерывных переменных. Если диапазоны предикторов неизвестны, неясно, к кому может быть применима модель. Например, применение предсказательной модели, разработанной на данных участников в возрасте от 30 до

60 лет, к лицам в возрасте 65 лет является экстраполяцией [186].

Многочисленные систематические обзоры показали, что отчёты исследований часто содержат недостаточно информации для проверки или применения модели у других лиц [43, 62, 66, 88]. Например, в достаточном для этих целей объёме представлена информация только в 13 из 54 исследований (24%), посвящённых разработке моделей прогнозирования рака молочной железы [43], и в 22 из 41 модели (54%) предсказания смерти для глубоко недоношенных младенцев [66]. В другом обзоре сообщалось, что ни в одном из включённых исследований не были представлены диапазоны непрерывных переменных [53], а в двух недавних систематических обзорах обнаружено, что даже возрастные диапазоны часто не указывались [73, 74].

Пункт 15б. Объясните, как использовать предсказательную модель (P).

Примеры

См. **табл. 15–17** и **рис. 5–7**.

Пояснение

Авторы должны объяснить, как разработанная модель может быть использована для получения вероятности предсказываемого исхода или рисков для отдельных лиц. Регрессионные модели определяют линейный предиктор (*linear predictor*) — взвешенную сумму значений предикторов модели (в виде определённой величины или кодов), где веса — коэффициенты регрессии (пункт 15а). При прогнозировании линейный предиктор часто называют прогностическим индексом (*prognostic index*). Коэффициенты регрессии логистического регрессионного анализа представляют собой логарифмы отношения шансов (*odds ratios*), в моделях Кокса — логарифмы отношения рисков (*log hazard ratios*). Регрессионные модели, за исключением моделей Кокса для данных типа «время до события», также включают свободный коэффициент (константу).

Предсказываемая вероятность исхода может быть оценена любой комбинацией значений предикторов. В логистической регрессионной модели её вычисляют следующим образом:

$$\text{Вероятность}=\exp(\beta_1 X_1 + \beta_x X_2 + \ldots + \beta_k X_k)/$$
$$/(1 + \exp(\beta_1 X_1 + \beta_x X_2 + \ldots + \beta_k X_k))=1/$$
$$/(1 + \exp(-(\beta_1 X_1 + \beta_x X_2 + \ldots + \beta_k X_k))),$$

где β_j — коэффициент регрессии для предиктора X_j; β_0 — свободный коэффициент. Это может помочь читателям представить вычисление более ясно. Последующее умножение на 100 преобразует вероятность в процент риска (% риска=100 × вероятность).

Для прогностических моделей, основанных на регрессии Кокса, предсказываемая вероятность исхода, наступающего в определённое время *t*, рассчитывается на основе прогностического индекса и оценки

Таблица 15. Пример таблицы. Полная модель, исходная вероятность события (baseline survival) в определённый момент времени и данные гипотетического человека для иллюстрации схемы индивидуального предсказания

(Упрощённая) модель B, шкала риска Reynolds

10-летний риск развития сердечно-сосудистых заболеваний (%)=[1 — $0{,}98634^{(\exp[B-22{,}325])}$] × 100%, где B=0,0799 × возраст + + 3,137 × натуральный логарифм (систолическое артериальное давление) + 0,180 × натуральный логарифм (высокочувствительный С-реактивный белок) + 1,382 × натуральный логарифм (общий холестерин) — 1,72 × натуральный логарифм (холестерин липопротеинов высокой плотности) + 0,134 × гемоглобин A_{1c} (%) (при диабете) + 0,818 (если курит в настоящее время) + 0,438 (если есть семейные случаи преждевременного инфаркта миокарда)

Клинический пример: оценка 10-летнего риска для 50-летней курящей женщины без диабета согласно модели ATP III или клинической упрощённой модели B (шкала риска Reynolds)

				Клинические показатели		Оценка 10-летнего риска, %	
Артериальное давление, *мм рт. ст.*	Холестерин, *мг/дл**			вчСРБ, *мг/л*	Семейный анамнез†	Модель ATP III	Упрощённая модель B
	Общий	ЛПВП	Не-ЛПВП				
155/85	240	35	205	0,1	Нет	11,5	4,9
155/85	240	35	205	0,5	Нет	11,5	6,5
155/85	240	35	205	1,0	Нет	11,5	7,4
155/85	240	35	205	3,0	Нет	11,5	8,9
155/85	240	35	205	5,0	Нет	11,5	9,7
155/85	240	35	205	8,0	Нет	11,5	10,5
155/85	240	35	205	10,0	Нет	11,5	10,9
155/85	240	35	205	20,0	Нет	11,5	12,3

Примечание. ATP — Adult Treatment Panel; ЛПВП — липопротеины высокой плотности; вчСРБ — высокочувствительный С-реактивный белок. Заимствовано из источника [208]. * Для преобразования значений холестерина из мг/дл в ммоль/л умножьте на 0,0259. † Инфаркт миокарда у родителей в возрасте до 60 лет.

Таблица 16. Пример таблицы. Простая балльная система индивидуальной оценки рисков (вероятностей) исхода*
Разработали карту расчёта риска для отдельных сотрудников. Умножили коэффициенты регрессии на 4 и округлили до ближайшего целого числа, получив баллы для каждого предиктора. Положительные оценки предикторов суммировали для подсчёта общей суммы баллов. Общий балл соответствует риску отпуска по болезни во время последующего наблюдения.

Отпуск по болезни в течение предыдущих двух месяцев		
Нет	0	. . .
0–1 неделя	2	. . .
>1 недели	3	. . .
Интенсивность боли в плече (0–10)		
0–3 баллов	0	. . .
4–6 баллов	2	. . .
7–10 баллов	3	. . .
Предполагаемая причина: растяжение или перенапряжение во время регулярной деятельности	3	. . .
Психологические проблемы, о которых сообщали сотрудники (тревожность, подавленное состояние, депрессия)	6	. . .
		___+
Общий балл		. . .

Общий балл	**Риск**
⩽1	10–20%
2–3	20–30%
4–5	30–40%
6–7	40–50%
8	50–60%
9–10	60–70%
11–12	70–80%
13–15	80–90%

Примечание. Заимствовано из источника [407]. * Прогнозируемую вероятность отпуска по болезни в течение 6 месяцев определяли по следующей формуле: P=1/[1+ exp — (–1,72 + 0,53 × отпуск по болезни 0–1-я неделя + 0,77 × отпуск по болезни >1 недели + 0,50 × боль в плече (4 –6 баллов) + 0,65 × боль в плече (7–10 баллов) + 0,68 × перенапряжение в результате обычной деятельности + 1,38 × сопутствующие психологические проблемы)]. Инструкции: если предиктор получил положительную оценку, необходимо указать его значение. Полученные баллы суммируются, чтобы получить общий балл. С помощью таблицы, следующей за картой баллов, можно рассчитать риск (%) отпуска по болезни для отдельного пациента на основе значений общего балла.

Таблица 17. Пример таблицы. Описание расчёта предсказываемой вероятности для отдельных лиц

Полученная логит-модель после подгонки к обучающим данным может быть выражена следующим образом:

$$\log\left(\frac{P_{\text{успешн.}}}{1-P_{\text{успешн.}}}\right) = 2{,}66 + 1{,}48*\text{неп.выкидыш} - 1{,}63*\text{миним.кровотечение} - 0{,}07*\text{возраст},$$

где $P_{\text{успешн.}}$ означает вероятность для пациента получить преимущество от выжидательной тактики. Переменной «*неп.выкидыш*» присваивается значение 1 в случае неполного выкидыша, диагностированного при первичном обследовании, и 0 — в иных случаях. Показателю «*мин.кровотечение*» присваивается значение 1 при отсутствии вагинального кровотечения или кровяных сгустков, 0 — в иных случаях.
В качестве альтернативного варианта модель для определения вероятности успешной выжидательной тактики ведения пациента можно представить в следующем виде:

$$P_{\text{успешн.}} = \frac{e^{2.66 + 1.48\text{IncompMisc} - 1.63\text{NilBleeding} - 0.07\text{Age}}}{1 + e^{2.66 + 1.48\text{IncompMisc} - 1.63\text{NilBleeding} - 0.07\text{Age}}}$$

Примечание. Информация из источника [409].

исходной выживаемости $S_0(t)$ [112, 274, 411]. Вычисление вероятности выживаемости производится по формуле:

$$S_0(t)^{\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)},$$

а вероятности исхода — по формуле:

$$1 - S_0(t)^{\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}.$$

Преобразование вероятности в проценты может быть выполнено умножением на 100.

Исследования, в которых разрабатываются новые предсказательные модели, часто направлены на создание простой в использовании модели или системы балльной оценки, часто называемой экспресс-моделью (*bedside model*) [45, 53]. Под простой моделью или системой балльной оценки мы подразумеваем упрощение формата представления базовой регрессионной модели, а не сокращение модели путём использования меньшего количества предикторов (как описано в пункте 10б). Многие известные предсказательные модели преобразованы или упрощены для облегчения их использования на практике, например, модель SCORE для прогнозирования 10-летнего риска смерти от сердечно-сосудистых заболеваний [140] или модель Уэллса (*Wells model*) для диагностики эмболии лёгочной артерии [412]. Упростить модель можно несколькими способами, например, путём преобразования (округления [413]) коэффициентов регрессии для каждого предиктора окончательной модели в легко суммируемые целые числа, которые затем соотносятся

с вероятностями исхода или выживаемости, как показано в приведённых выше примерах [414]. Крайняя форма округления коэффициентов регрессии состоит в том, чтобы присвоить каждому предиктору в окончательной модели одинаковый вес и просто подсчитать количество имеющихся факторов риска. Эти простые для суммирования баллы и соответствующие им вероятности исхода можно представить в виде таблиц или графиков, как показано выше.

Любое упрощение разработанной предсказательной модели путём округления чисел приведёт к некоторой потере точности предсказания [1, 413]. Следовательно, если авторы преобразуют формулу исходной модели в упрощённое правило оценивания, полезно представить показатели точности предсказания (например, *c*-индекс) (пункты 10г и 16) до и после упрощения. Эти сведения помогут читателям понять, в какой степени использование упрощённой модели приводит к потере точности предсказания. При этом упрощённая оценка должна основываться на исходной шкале коэффициентов регрессии (т.е. на логарифмической шкале шансов или рисков), а не на каком-либо преобразовании этих коэффициентов, таких как отношения шансов или рисков [415]. В частности, необходимо тщательно продумать, каким образом присваивать баллы предикторам, у которых соответствующее отношение шансов или рисков равно 1 или меньше (т.е. с нулевым или защитным/негативным эффектом на исход). В таких случаях присвоение положительного балла фактически увеличивает общий балл, что будет указывать на более высокую вероятность возникновения заболевания, тогда как соответствующий вклад должен быть ниже.

Если разрабатывается упрощённая система оценки риска, авторы должны ясно и однозначно описать шаги, предпринятые для создания упрощённой модели, и связь полученных баллов с вероятностью исхода. Система оценки риска может быть представлена в виде таблицы или карты с указанием возможных баллов и связанных с ними вероятностей исхода. Значения упрощённой модели могут быть сгруппированы для создания групп риска или вероятностей (пункт 11). В этом случае все участники со значениями модели в определённом диапазоне относятся к одной и той же группе риска и, таким образом, всем присваивается одинаковый риск. Однако простая констатация того, что участник, например, входит в группу низкого, промежуточного или высокого риска, без количественной оценки фактического предсказываемого риска, ассоциированного с каждой из групп, будет неинформативной. Группы, определённые на основании балльной оценки, должны быть связаны с соответствующими (средними или диапазоном) вероятностями исхода, представленными наблюдаемыми или предсказываемыми значениями рисков, либо и тем и другим.

Для моделей выживаемости кривые Каплана–Мейера должны быть представлены для каждой группы риска, поскольку с их помощью можно наглядно продемонстрировать вариации прогноза (например, дискриминационную способность модели). Кривые Каплана–Мейера можно дополнить общим количеством пациентов, количеством пациентов с исходом и описанием времени до события (с доверительными интервалами) для каждой группы.

Предсказательные модели иногда представляют в виде номограмм [310]. Этот формат представления предназначен не для упрощения разработанной модели, а для графической иллюстрации оригинальной математической формулы регрессии [112]. Такой формат может быть не знаком многим читателям и потенциальным пользователям, поэтому важно дать чёткие инструкции, как использовать номограмму для получения индивидуального предсказания. Номограмма не заменяет полное описание уравнения регрессии (пункт 15а).

Наконец, представление клинических сценариев и рабочего примера применения предсказательной модели к гипотетическому индивидууму с определённым профилем предикторов может быть поучительным независимо от способа представления самой модели.

Эффективность модели

Пункт 16. Сообщите показатели эффективности (включая доверительные интервалы) предсказательной модели (Р; П).

Примеры

См. **рис. 8–10** и **табл. 18**.

Пояснение

Все показатели эффективности, описанные в разделе «Методы» (пункт 10г), должны быть представлены и в разделе «Результаты», предпочтительно с указанием доверительных интервалов. В случае разработки или проверки нескольких моделей необходимо указать показатели эффективности для каждой из них. Для исследований по разработке моделей следует сообщать результаты внутренней проверки, включая любые показатели эффективности с поправкой на чрезмерный оптимизм (например, фактический и скорректированный *c*-индекс) (пункт 10б и **вставка Е**). Если предсказательная модель была упрощена (пункт 15б), следует описать эффективность (например, *c*-индекс) как оригинальной (например, представить полную регрессионную модель), так и упрощённой модели.

Помимо количественной оценки, рекомендуется визуализировать различия характеристик предсказания между лицами с исходом и без него в графическом виде с использованием гистограммы, графика плотности распределения (*density plot*), точечной диаграммы (*dot plot*) [417]. Для логистических регрессионных моделей можно представить характеристическую (ROC) кривую (*a receiver-operating characteristic curve*), но при условии чёткого отображения на ней предсказываемых рисков. В противном случае этот график будет малоинформативным и ничего не даст в дополнение к *c*-индексу.

Женщины

Фактор риска	Дополнительные баллы к шкале риска											Шкала риска
Возраст (годы)	35–39	40–44	45–49	50–54	55–59	60–64	65–69	70–74				
Дополнительные баллы для курящих	0 +13	+5 +12	+9 +11	+14 +10	+18 +10	+23 +9	+27 +9	+32 +8				
Систолическое артериальное давление (мм рт. ст.)	110–119 0	120–129 +1	130–139 +2	140–149 +3	150–159 +4	160–169 +5	170-179 +6	180–189 +8	190–199 +9	200–209 +10	>210 +11	
Общий холестерин (ммоль/л)	<5 0	5.0–5.9 0		6.0–6.9 +1		7.0–7.9 +1		8.0–8.9 +2			>9 +2	
Рост (м)	<1.45 +6	1.45–1.54 +4		1.55–1.64 +3		1.65–1.74 +2		>1.75 0				
Креатинин (мкмоль/л)	<50 0	50–59 +1	60–69 +1	70–79 +2	80–89 +2	90–99 +3	100–109 +3	>110 +4				
Инфаркт миокарда в анамнезе	Нет 0	Да +8										
Инсульт в анамнезе	Нет 0	Да +8										
Гипертрофия левого желудочка	Нет 0	Да +3										
Диабет	Нет 0	Да +9										
Общая оценка риска* =												

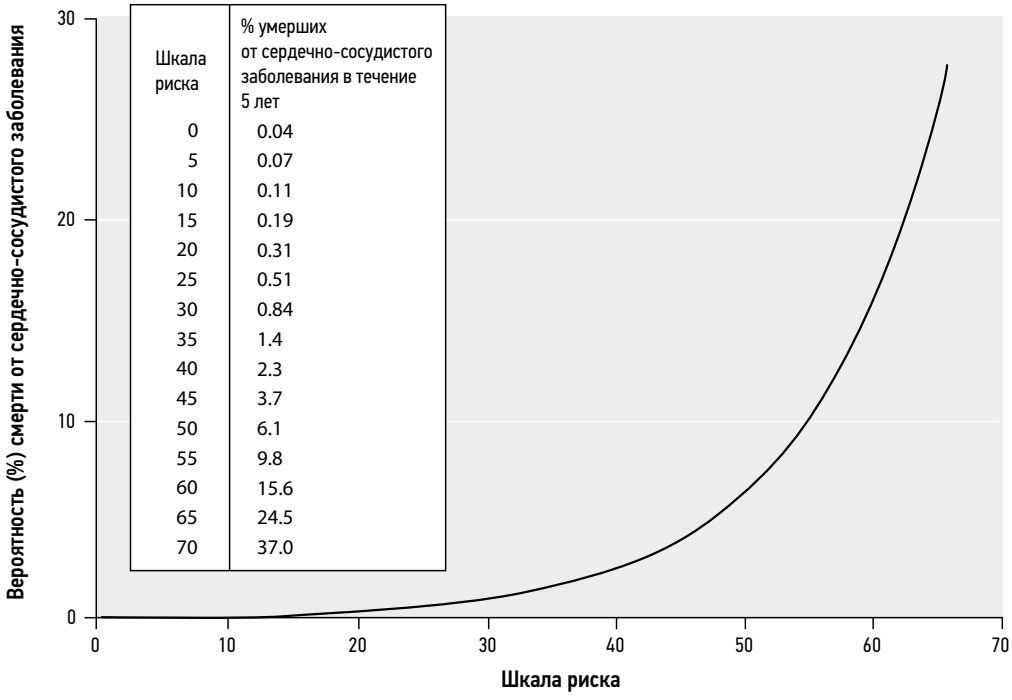
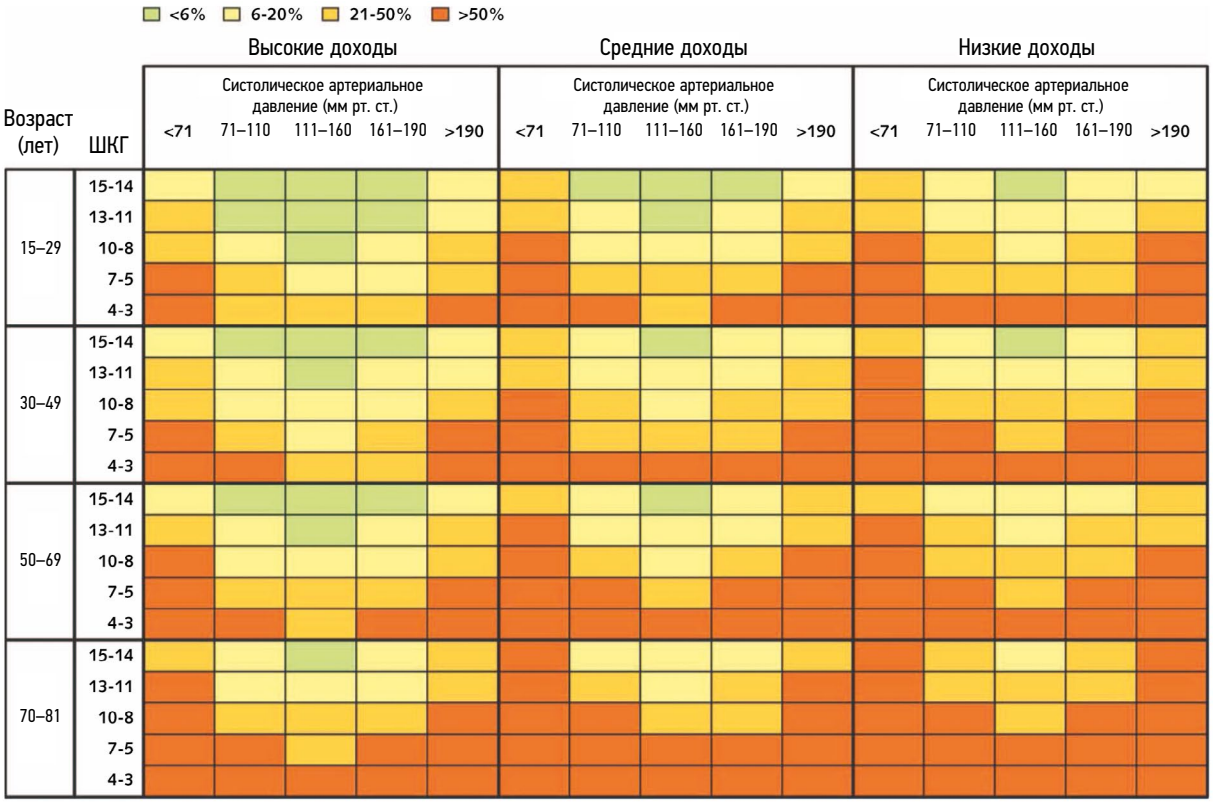


Рис. 5. Пример рисунка. Иллюстрированная система подсчёта баллов для определения предсказываемых вероятностей у отдельных лиц.

Примечание. Воспроизведено из источника [408] с разрешения BMJ Publishing Group.



Мы разработали простую прогностическую модель для использования в местах оказания медицинской помощи. В неё включили самые сильные предикторы с теми же квадратичными и кубическими членами, которые присутствовали в полной модели, но с поправкой на введение транексамовой кислоты. Представили прогностическую модель в виде диаграммы, где эти предикторы представлены несколькими категориями и перекрёстно табулируются. Категории были определены с учётом клинических и статистических критериев. Каждая ячейка диаграммы соответствует оценке риска для человека со значениями каждого предиктора в середине диапазона значений предиктора для этой ячейки. Ячейки диаграммы раскрашены в 4 цвета в соответствии с диапазонами вероятности смерти: <6%, 6–20%, 21–50%, >50%. Пороговые значения для этих диапазонов определили, исходя из отзывов потенциальных пользователей этой прогностической модели и предыдущих публикаций. ШКГ — шкала комы Глазго. Воспроизведено из источника [123] с разрешения BMJ Publishing Group.

Рис. 6. Пример рисунка. Графическая схема подсчёта баллов для определения предсказываемых вероятностей у отдельных лиц

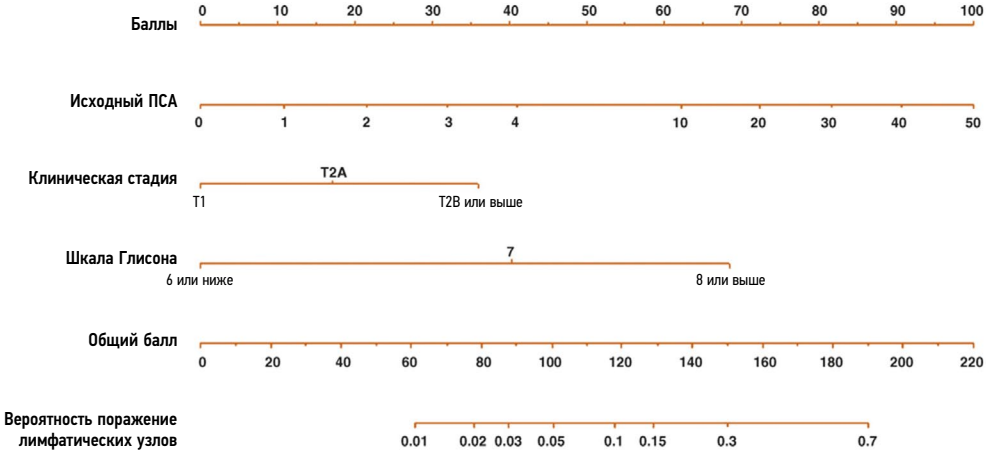


Рис. 7. Пример рисунка. Номограмма и её использование для индивидуального предсказания вероятности.

Примечание. Номограмма для прогнозирования поражённых лимфатических узлов у пациентов, перенёсших стандартную тазовую лимфатическую диссекцию. Инструкции: найдите значение простат-специфического антигена (ПСА) пациента, определённое до начала лечения, на оси исходных значений ПСА. Проведите линию прямо вверх к оси «Баллы», чтобы определить количество баллов, отражающих вероятность поражения лимфатических узлов, которые получит пациент исходя из его значений ПСА. Проделайте то же самое для каждой переменной. Суммируйте баллы, полученные для каждого предиктора. Найдите значение общей суммы на оси «Общий балл». Проведите линию прямо вниз, чтобы определить вероятность наличия у пациента поражённых лимфатических узлов. IPSA — initial prostate-specific antigen, LNI (lymph node involvement) — поражение лимфатических узлов. Перепечатано из источника [410] с разрешения Elsevier.

Примеч. ред. В оригинальном тексте руководства TRIPOD приведён рисунок, не подходящий под представленное в Примечании к этому рисунку описание. В настоящем переводе он заменен на рисунок, заимствованный из цитируемого авторами руководства источника [410].

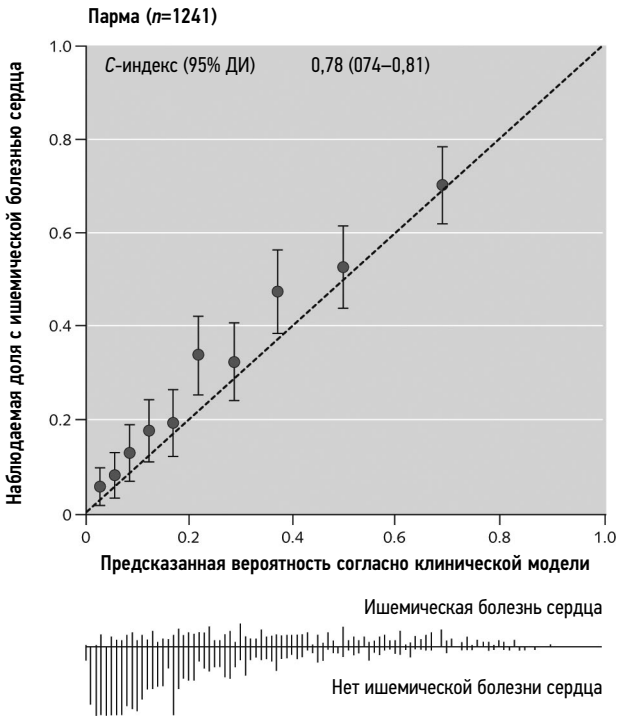


Рис. 8. Пример рисунка. Калибровочная кривая с с-индексом и распределением предсказываемых вероятностей для отдельных лиц с исходом (ишемическая болезнь сердца) и без него.

Примечание. Воспроизведено из источника [256] с разрешения BMJ Publishing Group.

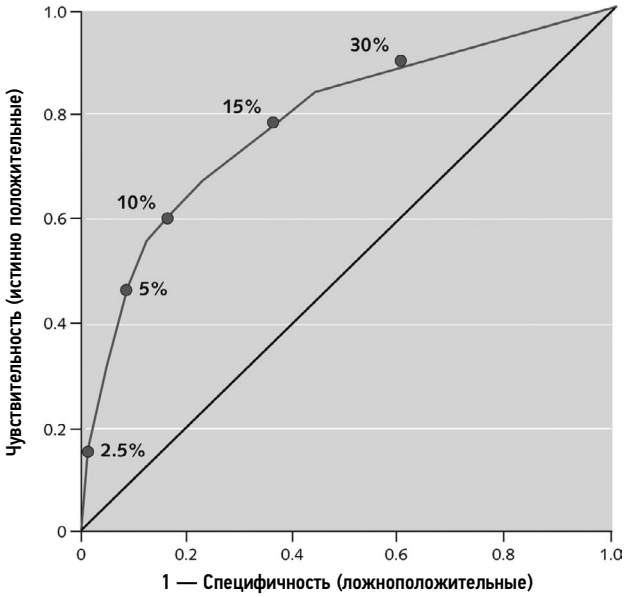


Рис. 9. Пример рисунка. Характеристическая кривая с отметками предсказываемых рисков.

Примечание. Характеристическая кривая для риска пневмонии... На графике показана чувствительность и специфичность нескольких пороговых значений риска предсказательной модели. Воспроизведено из источника [416] с разрешения BMJ Publishing Group.

Если разработано несколько моделей (например, базовая и расширенная) [418]), или если их оценивали на основе одного и того же набора данных, то можно сравнить их эффективность с помощью статистического метода, учитывающего тот факт, что модели были разработаны или проверены на тех же данных [334, 335, 419].

Если был рассчитан NRI (индекс реклассификации, *net reclassification improvement* — сумма долей правильно реклассифицированных наблюдений с наступившим и ненаступившим исходом. — *Примеч. ред.*), то принимая во внимание предостережения, описанные в пункте 10г, для оценки пользы добавления к существующей модели нового предиктора, авторы должны представить компоненты индекса, рассчитанные для наблюдений наступившего и ненаступившего исхода [339], а также суммарный показатель [351, 357, 420, 421].

Аналитические показатели, используемые для принятия решений, такие как чистая выгода (*net benefit*) или относительная полезность (*relative utility*), обычно представляют графически, а не в виде одной числовой оценки [361–363, 422]. Ось x в таких графиках представляет предпочтения пациента или клинициста (например, минимальную вероятность рака, при которой пациент выберет биопсию [138], или число пациентов, которых врач готов лечить, чтобы предотвратить одно сердечно-сосудистое событие [117]). Диапазон оси x в общем случае следует выбирать так, чтобы он представлял разброс значений показателя, наблюдаемый в обычных условиях. Например, кажется неоправданным включение 80% в качестве порогового значения для принятия решения о проведении биопсии для обнаружения рака простаты, поскольку в этом случае предполагается, что некоторые пациенты откажутся от биопсии при 75%-й вероятности рака.

Ось *y* (чистая выгода) отображает разницу между количеством истинно положительных и ложноположительных результатов, взвешенную с учётом коэффициента, который даёт стоимость ложноположительного результата по сравнению с ложноотрицательным, например (**рис. 10**). В оригинальном тексте руководства TRIPOD указан рисунок под номером 18, однако в руководстве нет рисунка с таким номером. По смыслу, речь идет о рисунке № 10. — *Примеч. ред.*), если две модели, сравниваемые при определённом пороговом значении, имеют разницу по показателю чистой выгоды 0,005 [т.е. модель A (QRISK2-2011) минус модель B (NICE Framingham)], это означает чистое увеличение истинно положительных результатов, т.е. при использовании модели A выявляется ещё 5 истинно положительных результатов на 1000 человек без увеличения количества ложноположительных результатов.

При графическом представлении аналитических показателей принятия решений не следует отводить бóльшую часть графика потерям чистой выгоды (*negative net benefit*). Кривые должны быть сглажены; если размер выборки небольшой, исследователи могут применить либо метод статистического сглаживания, либо рассчитать чистую выгоду с более широкими интервалами (например, каждые 5%).

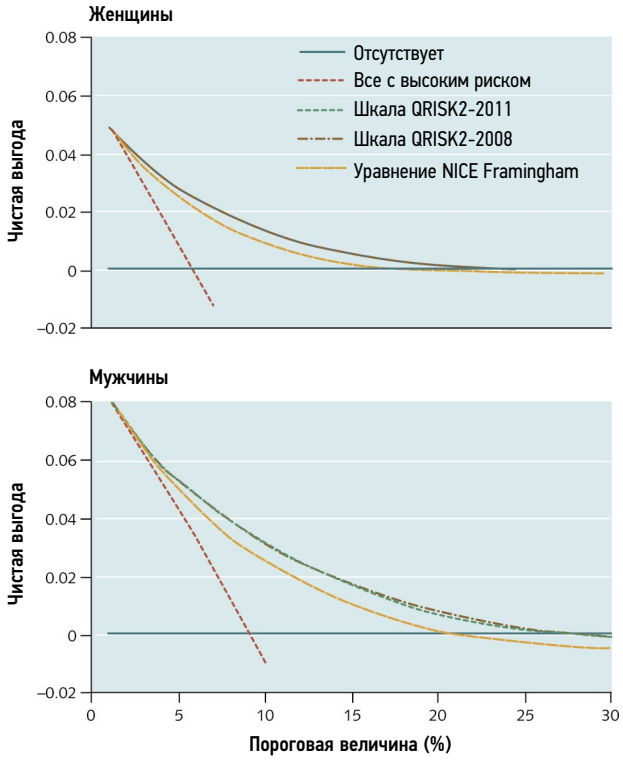


Рис. 10. Пример рисунка. Анализ кривой принятия решений.

Примечание. На рисунке показаны кривые чистой выгоды для шкал QRISK2-2011, QRISK2-2008 и уравнения NICE Framingham, применимых к лицам в возрасте от 35 до 74 лет. При традиционном пороговом значении 20%, используемом для обозначения лиц с высоким риском развития сердечно-сосудистых заболеваний, чистая выгода от использования шкалы QRISK2-2011 для мужчин, по сравнению с уравнением NICE Framingham, заключается в выявлении дополнительных 5 случаев на 1000 человек без увеличения количества пациентов, получающих лечение без необходимости. Чистая выгода от использования шкалы QRISK2-2011 с пороговым значением 20% для женщин заключается в выявлении двух дополнительных случаев на 1000 человек, по сравнению с неиспользованием какой-либо модели (или уравнения NICE Framingham). По-видимому, нет чистой выгоды от использования 20%-го порога для уравнения NICE Framingham для выявления женщин с повышенным риском развития сердечно-сосудистых заболеваний в течение следующих 10 лет. NICE — National Institute for Health and Care Excellence. Воспроизведено из источника [117] с разрешения BMJ Publishing Group.

Таблица 18. Пример таблицы реклассификации [с индексом реклассификации (*Net Reclassification Improvement*) и 95% ДИ] для базовой и расширенной моделей с использованием одинакового порога вероятности*

ТГВ присутствует (*n*=416)			
	Модель 2 с D-димером		
Модель 1 без D-димера	⩽25	>25	Всего
⩽25	92	123	215
>25	26	175	201
Всего	118	298	416
ТГВ отсутствует (*n*=1670)			
	Модель 2 с D-димером		
Модель 1 без D-димера	⩽25	>25	Всего
⩽25	1223	116	1339
>25	227	104	331
Всего	1450	220	1670

Примечание. ТГВ — тромбоз глубоких вен. Заимствовано из источника [367]. * Увеличение индекса реклассификации при добавлении теста D-димера к комбинации данных анамнеза и осмотра с использованием показанных значений... составило: (0,30 — 0,06) — (0,07–0,14)=0,31 (95% ДИ 0,24– 0,36).

Обновление модели

Пункт 17. Если применимо, сообщите результаты любого обновления модели (т.е. состава модели, условий её применения, характеристик эффективности) (П).

«Для повторно откалиброванных моделей все коэффициенты регрессии умножили на угловой коэффициент (*slope*) калибровочной модели (0,65 для мужчин и 0,63 для женщин). Свободный коэффициент (*intercept*) был скорректирован путём умножения исходного значения на калибровочный угловой коэффициент (*calibration slope*) и добавления соответствующего свободного коэффициента калибровочной модели (–0,66 для мужчин и –0,36 для женщин). С целью обновления моделей дополнительно скорректировали коэффициенты регрессии предикторов, имевших дополнительную ценность для повторно откалиброванной модели. Для мужчин коэффициенты регрессии дополнительно корректировали с учётом предикторов отстранения (от донорства. — *Примеч. ред.*) во время предыдущего визита, времени с предыдущего визита, разницы концентраций гемоглобина (между визитами. — *Примеч. ред.*) и фактора сезонности, для женщин — с учётом отстранения (от донорства. — *Примеч. ред.*) во время предыдущего визита и разницы концентраций гемоглобина... доступны в качестве вспомогательной информации в онлайн-версии настоящей статьи для построения точных формул откалиброванных и пересмотренных моделей с целью оценки риска отстранения от донорства на основании значений концентрации гемоглобина [370].» (Диагностика; Обновление модели; Логистический.)

«Ошибки в калибровке, допущенные при применении первого подхода, послужили причиной повторной калибровки. Мы получили равномерный коэффициент сжатия (*uniform shrinkage factor*), когда использовали logit(P(Y=1))=a + b*logit(p) для второго подхода. Мы получили оценки a=–1,20 и b=0,11, что свидетельствует о сильном сжатии [368]». (Диагностика; Обновление модели; Логистический.)

«Показатели эффективности оригинальной клинической предсказательной модели, по сравнению с другими моделями, расширенными за счёт включения генетических переменных, отобранных методом лассо, представлены в **табл. 3**. Для проверки согласованности двух моделей выполняли тесты отношения правдоподобия. Площадь под кривой (AUC) в оригинальной клинической модели составляла 0,856. Добавление ОНП (однонуклеотидные полиморфизмы) TLR4 к клинической модели

привело к небольшому уменьшению AUC, а добавление данных о полиморфизме TLR9-1237 — к небольшому увеличению AUC до 0,861, хотя это изменение не было значимым (p=0,570). Использование данных о полиморфизмах NOD2 также не улучшили модель [423]». (Прогнозирование; Обновление модели; Логистический.)

Пояснение

Эффективность существующей модели на новом наборе данных часто ниже, чем на оригинальных данных, использованных для разработки этой модели. По этой причине исследователи могут обновлять или повторно калибровать существующую модель одним из нескольких способов, описанных выше (**табл. 3** и пункт 10д). Если предсказательная модель была обновлена, исходя из результатов проверочного исследования, то авторы должны сообщить обо всех аспектах модели, которые были изменены. В зависимости от метода обновления (**табл. 3**) это могут быть сведения о пересчитанном свободном коэффициенте, обновлённых коэффициентах регрессии (например, с использованием углового коэффициента калибровочного графика оригинальной модели, определённого в проверочном наборе данных) или установленных коэффициентах регрессии модели при добавлении новых предикторов. Обновление модели в контексте регрессионных моделей Кокса является более сложным процессом [309, 373].

Обновлённая модель — это, по сути, новая модель. Поэтому обновлённые модели должны быть представлены также достаточно подробно, чтобы читатели могли делать предсказания для отдельных пациентов (пункты 15a и 15b) в последующих проверочных исследованиях или на практике. Также необходимо указать все показатели эффективности обновлённых моделей (пункт 16).

Обсуждение

Ограничения

Пункт 18. Обсудите любые ограничения исследования (например, нерепрезентативная выборка, недостаточное количество событий на один предиктор, отсутствующие данные) (Р; П).

Примеры

«Самое важное ограничение модели для предсказания длительного пребывания в отделении интенсивной терапии (ОИТ) — это её сложность. Мы полагаем, что сложность обусловлена большим количеством факторов, определяющих длительность пребывания в ОИТ. Эта сложность требует использования автоматизированного сбора данных и выполнения необходимых расчётов. С учётом ограниченного доступа большинства медицинских учреждений к современным информационным технологиям это существенное препятствие к широкому применению модели на практике. По мере того как всё больше учреждений включают электронные медицинские карты в свой процесс, модели, подобные описанной здесь, могут иметь бóльшую ценность.

Наши результаты имеют ряд дополнительных ограничений. Во-первых, полезность модели, вероятно, ограничена территорией США из-за международных различий, влияющих на пребывание в ОИТ. Эти различия, вероятно, негативно повлияют на использование пятых суток пребывания в ОИТ как порога для дальнейшего нахождения в отделении. Во-вторых, сбор физиологических данных в первые сутки не учитывает влияние осложнений и ответа на терапию, хотя оценка эффектов этих явлений даже на пятые сутки также может быть преждевременной. Предыдущие исследования показали, что более половины осложнений в ОИТ возникают после пятых суток пребывания в отделении. В-третьих, при всей своей сложности модель не учитывает дополнительные факторы, которые, как известно, влияют на длительность пребывания в ОИТ. К ним относятся нозокомиальная инфекция, отказ от реанимации, укомплектованность ОИТ врачами, паралич в ОИТ, практика седации в ОИТ. В-четвёртых, самый бóльшой недостаток модели — неточность в предсказании (до двух суток) оставшегося времени пребывания в ОИТ. Мы предполагаем, что такая неточность объясняется задержкой выписки, чтобы избежать перевода в ночное время или выходные дни, а также частотой осложнений, возникающих на 6–8-е сутки пребывания в отделении» [424]. (Прогнозирование; Разработка; Проверка.)

«Эта работа имеет ряд ограничений. Во-первых, оценки успеваемости резидентов по одной программе с единственной специальностью. Кроме того, в нашей программе рассмотрена лишь небольшая часть общей популяции студентов-медиков в США. Воспроизводимость наших результатов в других условиях и программах неизвестна. Во-вторых, для оценки успеваемости резидентов мы использовали субъективные общие критерии в сочетании с итоговыми оценками. Хотя межэкспертная надёжность (*interrater reliability*) в нашем исследовании была высокой, золотого стандарта для оценки клинической успеваемости не существует, и лучший метод для этого остаётся предметом дискуссии. Наконец, показатель r^2=0,22 в нашем регрессионном анализе показывает, что бóльшая часть дисперсии средних оценок успеваемости осталась необъяснённой. Это может быть связано с ограниченной информацией в заявках на резидентуру о таких важных сферах, как лидерские качества, способность работать в команде и профессионализм» [425]. (Прогнозирование; Разработка.)

Пояснение

Даже самые лучшие исследования предсказательных моделей, вероятно, будут иметь множество ограничений, которые необходимо учитывать. Тем не менее во многих статьях, опубликованных даже в самых влиятельных журналах, не сообщается об ограничениях [426]. Более того, в исследованиях молекулярных диагностических

маркеров часто можно наблюдать чрезмерно оптимистичную интерпретацию результатов без должного учёта ограничений, вытекающих из дизайна и результатов исследования [158].

После публикации многие соавторы статьи отмечают, что напечатанное обсуждение не полностью отражает их точку зрения и не содержит ограничений и предостережений [427]. Тем не менее неоднократно утверждалось, что явное признание ограничений является одним из ключевых аспектов научной работы и наиболее ценной частью обсуждения научной статьи [428, 429]. Признание ограничений усиливает, а не ослабляет исследование.

Ограничения необходимо рассматривать в перспективе, и следует приложить усилия, чтобы охарактеризовать влияние, которое может оказать каждая отдельная проблема на результаты исследования. В некоторых случаях такое влияние может быть слишком неопределённым, и его последствия практически невозможно оценить. В других ситуациях направление систематической ошибки можно с уверенностью предсказать, а её эффект — достоверно оценить.

Ограничения могут относиться к любому аспекту дизайна исследования, его проведения или анализа данных. Они могут быть обусловлены [430], но не ограничиваться типами исследуемых популяций, выбором участников (репрезентативность), выбором предикторов, надёжностью определений и процедур, используемых при сборе данных о предикторах и исходах, размером выборки (особенно в сравнении со сложностью и количеством исследуемых предикторов и исходов), длительностью последующего наблюдения и методами регистрации исходов, многочисленностью анализов, отсутствующими данными, переобучением (*overfitting*), особенностями внутренней проверки модели и разницей между когортами для разработки и проверки модели (если применимо). Следует обсудить, повлияли ли ограничения на разработку модели и (или) результаты её проверки, и каким может быть их общее влияние на достоверность (*credibility*), применимость (*applicability*) и обобщаемость (*generalizability*) многофакторной модели.

Например, если в исследовании не учитывались хорошо известные предикторы, об этом необходимо сообщить и перечислить их. При этом следует уточнить, нужно ли учитывать эти предикторы в будущих исследованиях или на практике, или же включённые в модель предикторы содержат достаточно информации от пропущенных предикторов. Если же авторы предполагают, что полученные оценки будут неточными (чрезмерно оптимистичными), об этом также необходимо сообщить и уточнить, насколько серьёзными будут связанные с этим проблемы, насколько завышенными будут показатели эффективности модели; должно ли это повлиять на решение о дальнейшем применении модели на практике или потребуется отсрочка для её последующей проверки, обновления (включая повторную калибровку; пункты 10д и 17) или реализации стратегии непрерывного усовершенствования (например, шкала QRISK2 [117, 431–433], что сняло бы эти опасения.

В работах, в которых модель разрабатывается для одной популяции без какой-либо проверки в другой, отсутствие внешней проверки следует упомянуть по умолчанию как серьёзное ограничение, помимо любых других ограничений, которые могут существовать.

Интерпретация результатов

Пункт 19а. В случае проверочного исследования обсудите полученные результаты с упоминанием характеристик оригинальной модели, а также характеристик, полученных с использованием любых других проверочных данных (П).

Пример

«Шкала ABCD2 — результат совместных усилий команд под руководством Johnston и Rothwell, которые объединили два независимых набора данных для получения клинических сведений о высоком риске последующего инсульта. Набор данных команды Rothwell был небольшим, составлен из числа пациентов, направленных врачами первичного звена здравоохранения, и включал предикторные переменные, оценённые неврологом спустя 1–3 суток. Набор данных команды Johnston был получен из ретроспективного исследования с участием пациентов из Калифорнии, перенёсших транзиторную ишемическую атаку.

Последующие исследования шкалы ABCD2 либо были ретроспективными, либо исследованиями, в которых использовалась информация из баз данных. Ong и соавт. установили, что чувствительность шкалы в определении развития инсульта в течение семи дней составляет 96,6% для оценок более двух баллов, однако в этом случае в группу высокого риска были отнесены 83,6% пациентов. Fothergill и соавт. ретроспективно проанализировали регистр данных 284 пациентов и обнаружили, что при пороговом значении менее 4 баллов были пропущены 4 из 36 инсультов, наступивших в течение 7 суток. Asimos и соавт. ретроспективно рассчитали баллы по шкале ABCD2 на основе существующей базы данных, но при этом оценка риска не была выполнена для 37% пациентов, включая 154 из 373 пациентов, у которых в течение 7 последующих суток наступил инсульт. Sheehan и соавт. обнаружили, что шкала ABCD2 хорошо различает пациентов с транзиторной ишемической атакой или малым инсультом, по сравнению с пациентами с преходящими неврологическими нарушениями, вызванными другими состояниями, но они не оценивали предсказательную точность (*predictive accuracy)* оценки риска последующего инсульта. Tsivgoulis и соавт. поддержали использование 2 баллов по шкале ABCD2 как порогового для определения высокого риска на основании результатов небольшого проспективного исследования с участием

пациентов, которые перенесли транзиторную ишемическую атаку и были госпитализированы. Систематический обзор, выполненый Giles и Rothwell, показал, что объединенная оценка AUC составляет 0,72 (95% ДИ 0,63–0,82) для всех исследований, соответствующих критериям поиска, и 0,69 (95% ДИ 0,64–0,74) — после исключения исследований, в которых шкала была разработана. В нашем исследовании величина AUC находится в нижней части доверительного интервала этих результатов, приближаясь к 0,5» [434]. (Прогнозирование.)

Пояснение

Если в исследовании представлены результаты проверки существующей модели, авторы должны обсудить, идентична ли проверенная модель той, которая была разработана ранее, а если между ними есть какие-либо отличия, необходимо объяснить их (пункт 12). Должны быть обсуждены характеристики модели, зафиксированные в проверочном исследовании, в том числе в контексте характеристик этой модели, описанных в оригинальном исследовании, в котором модель была разработана. Следует выделить основные результаты, а также любые систематические ошибки, которые могли повлиять на результаты сравнения.

Если проверочное исследование демонстрирует другие (обычно худшие) показатели эффективности, следует обсудить причины. Например, ими могут быть различия характеристик групп исследований, определений или методов измерения предикторов и исходов, а также времени последующего наблюдения (если применимо). Если с использованием одного набора данных проверяют несколько моделей, т.е. проводят так называемую сравнительную проверку (*comparative validation*), следует опять же выделить основные результаты и указать на систематические ошибки, которые могли повлиять на результаты сравнения [47, 48].

Пункт 19б. Обсудите результаты с учётом целей, ограничений, результатов схожих исследований и других актуальных сведений (Р; П).

Пример

«Наши модели, основанные на демографических данных и лабораторных маркерах тяжести ХБП (хроническое заболевание почек. — *Примеч. авт.*), предназначены для предсказания риска развития почечной недостаточности. Подобно исследователям из Kaiser Permanente и группы исследования RENAAL, мы обнаружили, что более быстрое прогрессирование ХБП с развитием почечной недостаточности предсказывают более низкая расчётная СКФ (скорость клубочковой фильтрации. — *Примеч. авт.*), высокая альбуминурия, молодой возраст и мужской пол. Кроме того, более высокий риск почечной недостаточности предсказывают, а также повышают прогностическую значимость СКФ и альбуминурии низкие концентрации альбумина в сыворотке, кальция и бикарбоната, а также высокая концентрация фосфата в сыворотке. Эти маркеры позволяют более точно оценить фактическую СКФ или могут отражать нарушения канальцевой функции почек или лежащие в их основе воспаление или недостаточность питания.

Хотя связь этих лабораторных маркеров с прогрессированием ХБП уже была описана, мы объединили все имеющиеся сведения в едином уравнении риска (калькулятор риска и **табл. 5**, а также мобильное приложение доступны на сайте www.qxmd.com/Kidney-Failure-Risk-Equation). Кроме того, мы не обнаружили улучшения показателей эффективности модели при добавлении к ней анамнестических сведений (наличие диабета и гипертензии), а также результатов медицинского осмотра (систолическое и диастолическое артериальное давление, масса тела). Хотя эти переменные, несомненно, имеют важное значение для диагностики и лечения ХБП, отсутствие улучшений в эффективности модели можно объяснить высокой распространённостью этих состояний при ХБП и неточностями в определении тяжести заболевания после того, как уже были учтены расчётная СКФ и уровень альбуминурии» [261]. (Прогнозирование; Разработка; Проверка.)

Пояснение

Интерпретация результатов исследования помещает находки авторов в контекст других свидетельств — ранее проведённых схожих исследований той же многофакторной модели или разных моделей с тем же или схожим исходом или других свидетельств, которые могут считаться уместными. Обычно существует множество других предсказательных моделей, которые служат тем же или схожим целям. Например, только за один год опубликованы данные по 240 оценкам 118 различных инструментов предсказания одной только смертности [65].

При наличии множества доступных предсказательных моделей для одной целевой популяции или одинаковых исходов было бы полезно систематически сопоставлять имеющуюся модель с ранее разработанными для определения сильных и слабых сторон новой модели. В идеале такое сравнение выполняется на основе систематического обзора предыдущих исследований, если таковой проводился [47, 48, 435]. В противном случае авторам необходимо рассмотреть возможность проведения по крайней мере неформального обзора предыдущих свидетельств и обсудить основные исследования, которые могут конкурировать с текущей работой, с точки зрения убедительности полученных результатов и плана действий для дальнейших проверочных исследований или применения модели на практике. Также полезно прокомментировать отличия в построении моделей, изучаемых предикторах, целевой популяции и условиях применения модели, а также эффективности и значимости проверочного процесса. На интерпретацию результатов также могут влиять дополнительные соображения, в том числе ограничения исследования (пункт 18), были ли достигнуты первоначальные цели исследования и, если нет, то почему; а также

перспективы использования предложенной модели в различных условиях, ожидания от её внедрения в медицинскую практику.

В некоторых случаях может быть интересно рассмотреть другие уместные свидетельства. Например, это могут быть данные о биологической обоснованности (*biological plausibility*) использования предикторов, включённых в модель, или другие данные, которые могут дать представление о том, почему некоторые предикторы особенно важны для модели. Эмпирическое исследование показывает, что авторы, как правило, крайне непоследовательны в обсуждении биологических оснований, поддерживающих включение конкретных предикторов в модели [436]. Следует приложить усилия, чтобы высказать сбалансированное суждение и обсудить как поддерживающие, так и опровергающие свидетельства при наличии таковых.

Применение

Пункт 20. Обсудите потенциал клинического использования модели и значение для будущих исследований (Р; П).

Примеры

«Вероятность заболевания гриппом зависит от исходной вероятности возникновения гриппа в популяции, результатов клинического обследования и, возможно, результатов экспресс-тестов (*point of care tests*) для диагностики гриппа. Мы определяли вероятность заболевания гриппом в течение каждого сезона на основе данных Центров по контролю и профилактике заболеваний (Федеральное агентство, США. — *Примеч. ред.*). Недавний систематический обзор показал, что экспресс-тесты для диагностики сезонного гриппа имеют чувствительность примерно 72% и точность (*accuracy*) 96%. Используя эти данные о сезонной вероятности и точности тестов, при отношении правдоподобия (*likelihood ratios*) для оценки гриппа, равном 1, пороговом значении выполнения/невыполнения теста, равном 10%, и выполнения теста / начала лечения, равном 50%, мы обобщили применяемый подход к оценке пациентов с подозрением на грипп в **табл. 5**. В пик сезонной эпидемии гриппа врачи, желающие ограничить использование противогриппозных препаратов, должны рассмотреть целесообразность экспресс-тестирования даже у пациентов с высоким риском заболевания. Для пациентов с высоким риском осложнений необходимо рассмотреть необходимость проведения эмпирической терапии» [181]. (Диагностика; Разработка; Проверка; Клиническое использование.)

«Для дальнейшей оценки этих результатов необходимо решить ряд вопросов. Во-первых, хотя в исследование, из которого были получены данные, были включены амбулаторные пациенты, для этих анализов мы намеренно ограничили выборку исследования стационарными пациентами, так как частота случаев ПОТР (послеоперационная тошнота и рвота. — *Примеч. авт.*) среди амбулаторных пациентов была значительно ниже (34%), и потому что выполнялись различные типы хирургических вмешательств (например, не было операций на брюшной полости). Соответственно, наши результаты должны распространяться в первую очередь на стационарных больных. Следует отметить, что в настоящее время не существует правил, которые были разработаны как для стационарных, так и для амбулаторных пациентов. Это всё ещё предмет для будущих исследований, особенно с учётом увеличения объёмов амбулаторной хирургии» [437]. (Прогнозирование; Дополнительное значение; Клиническое использование.)

«Наше исследование имело несколько ограничений, которые следует признать. Мы объединили данные из двух разных популяций с несколько разными критериями включения, хотя итоговый набор данных имеет преимущество в большей обобщаемости (*generalizability*), поскольку он включает пациентов из двух стран, отобранных в течение двух разных эпидемиологических (по гриппу) сезонов и имеет общую претестовую вероятность, типичную для сезона гриппа. Кроме того, сбор данных был ограничен взрослыми, поэтому неясно, применимы ли эти результаты к пациентам младшего возраста. Несмотря на простоту, подсчёт баллов для оценки риска может быть слишком сложным для запоминания. В связи с этим может помочь программирование, реализованное в виде приложения для смартфонов или работы в сети Интернет» [181]. (Диагностика; Разработка; Проверка; Ограничения; Значение для будущих исследований.)

Пояснение

В разделе «Обсуждение» авторы могут и должны обсудить последствия проведённого исследования на нескольких уровнях. Предсказательные модели могут быть использованы с разными целями. В пункте 3а (актуальность и обоснование модели) исследователям предлагается описать их для своих моделей. «Обсуждение» — это тот раздел рукописи, где авторы могут обсудить потенциал клинического применения модели, исходя из полученных результатов исследования. Очевидно, что для недавно разработанных моделей может быть сложнее формально обсудить их применение на практике, поскольку следующим логическим шагом должно быть проведение проверочных исследований. Безусловно, авторам не следует рекомендовать применение модели, основываясь лишь на результатах первоначального исследования, в котором модель разрабатывалась.

Точно так же клинические руководства не должны рекомендовать использование непроверенных предсказательных моделей. Более того, клинические рекомендации должны быть основаны на наличии и синтезе свидетельств точности модели, проверенной на данных других участников и, следовательно, на воспроизводимости результатов модели в других условиях.

Следует подчеркнуть, что проверочные исследования на внешних (независимых) данных (*external model-validation studies*), даже проспективные, не показывают степень

влияния моделей на принятие медицинских решений или важные для здоровья исходы. Влияние на принятие решений, поведение врача и исходы пациентов можно оценить только в сравнительных (предпочтительно рандомизированных [438–440]), а не проверочных исследованиях с единственной когортой [20, 28, 33]. К сожалению, проверочные исследования на внешних данных проводятся редко, не говоря уже об исследованиях влияния моделей (*model-impact studies*) [441, 442].

Отвечая на вопрос о применимости результатов исследования, следует обсудить условия применения (учреждения первичной помощи, больницы), географическое положение, возраст, пол и клинические особенности медицинской проблемы, предсказание которой выполняется. Также следует уделить внимание тому, как (прогностическое) правило можно применить. Например, предназначена ли проверенная диагностическая модель для подтверждения или исключения заболевания, какие пороговые значения предсказательного правила могут быть использованы для достижения каждой цели и каковы возможные последствия применения модели (дальнейшие обследования, ложноположительные или ложноотрицательные результаты).

Помимо обсуждения возможных прямых последствий, авторы могут представить конкретные предложения по проведению дальнейших исследований с учётом ограничений настоящего исследования, уделяя внимание таким вопросам, как необходимость проверки новой модели в другом наборе данных, эффективность разработанного правила для достижения первоначально заявленных целей (включая потенциальную полезность других предикторов), выбор пороговых значений для определения клинической тактики, проблемы практического применения.

Другие сведения

Дополнительная информация

Пункт 21. Предоставьте информацию о доступности дополнительных материалов, таких как протокол исследования, веб-калькулятор и наборы данных (Р; П).

Примеры

«Дизайн и методы исследования RISK-PCI были ранее опубликованы [ссылка]. Вкратце, RISK-PCI — это наблюдательное (*observational*) продольное (*longitudinal*) когортное одноцентровое исследование, специально спланированное с целью разработки и проверки точной модели риска для предсказания основных неблагоприятных сердечно-сосудистых событий после ЧКВ (чрескожное коронарное вмешательство. — *Примеч. авт.*) у пациентов, предварительно получавших клопидогрел в дозе 600 мг. Пациенты были набраны в период с февраля 2006 г. по декабрь 2009 г. От каждого пациента получено информированное согласие. Протокол исследования соответствует этическим принципам Хельсинкской декларации. Он был одобрен локальным комитетом по этике исследований и зарегистрирован в Регистре текущих контролируемых исследований — ISRCTN83474650 (www.controlled-trials.com/ISRCTN83474650)» [443]. (Прогнозирование; Разработка.)

«Удобные в пользовании калькуляторы для оценки рисков по шкале Reynolds для мужчин и женщин имеются в свободном доступе на сайте www.reynoldsriskscore.org» [444]. (Прогнозирование; Дополнительное значение.)

«Открытые исходные коды для подсчёта балов по шкале QFracture доступны на сайте www.qfracture.org под лицензией GNU lesser general public licence, версия 3)» [315]. (Прогнозирование; Проверка.)

Пояснение

Все исследования с участием людей должны проводиться в соответствии с протоколом [445, 446]. Протокол исследования, проводимого с целью разработки предсказательной модели, должен начинаться с чётко сформулированной цели, за которой следует информация о дизайне исследования, описание предикторов и исхода, план статистического анализа. Исследования по разработке или проверке предсказательных моделей только выиграют от тщательного подготовленного и подробного протокола, составленного до начала проведения анализа. Такие протоколы периодически публикуются [447–464]. Сведения, изложенные в опубликованных протоколах, позволяют читателям сравнить то, что было запланировано, с тем, что было фактически сделано. Если протокол не был опубликован, мы рекомендуем авторам подавать протокол исследования в журнал вместе с рукописью и, по возможности, представить его вместе с опубликованной статьей в виде электронного (онлайн) приложения, что поможет рецензентам оценить опубликованный отчёт.

Для использования модели в повседневной практике или в дальнейших исследованиях необходимо сообщать достаточно подробные сведения о ней (пункты 15а, 15б и 16), чтобы можно было делать предсказания вероятности для отдельных лиц, а исследователям — проверять и обновлять предсказательную модель. Кроме того, авторам рекомендуется представить подробную информацию о том, как получить доступ к разработанным вебкалькуляторам и автономным приложениям, например, для электронных устройств, таких как планшеты (например, www.outcomes-umassmed.org/GRACE/). В редких случаях, когда описание предсказания является слишком сложным для включения со всеми подробностями в опубликованный отчёт (или в приложение), или если модель должна постоянно обновляться (например, QRISK2 [139]), необходимо представить подробную информацию о том, где можно получить полный доступ к исходному компьютерному коду для расчёта предсказания.

В последнее время растет понимание того, что наборы данных и компьютерные коды по возможности должны быть общедоступными. Это необходимо

для воспроизведения выполненного анализа [27, 465–467], а также для того, чтобы данные отдельных участников можно было объединить для метаанализа [468–473]. В помощь авторам было разработано руководство по подготовке сырых клинических данных (*raw clinical data*) и обмену ими с другими учёными [271]. В образцовом исследовании, проведённом Marchionni и соавт. [474], представлен прототип шаблона для воспроизводимой разработки прогностической модели, демонстрирующий возможность соблюдения принципа прозрачности всего процесса. Если возможно, авторы должны предоставить подробную информацию о доступе к исходному коду, используемому для анализа данных.

В настоящее время не существует обязательного требования о регистрации наблюдательных исследований. Эту идею поддержали многие [475–478], но были и те, кто возражал [479–481]. Во многих реегистрах клинических исследований, включая ClinicalTrials.gov, прямо указано, что наблюдательные исследования могут быть зарегистрированы [482]. Несмотря на очевидные трудности, связанные с подробным предварительным планированием некоторых типов наблюдательных исследований, (проспективные) исследования, в ходе которых собирают данные о новых участниках с целью разработки или проверки предсказательной модели, не должны вызывать подобных опасений [476].

Финансирование

Пункт 22. Укажите источник финансирования и роль спонсоров в настоящем исследовании (Р; П).

Примеры

«Проект Reynolds Risk Score был поддержан исследовательскими грантами от Donald W. Reynolds Foundation (Лас-Вегас, штат Невада). Дополнительное финансирование получено от Doris Duke Charitable Foundation (Нью-Йорк, штат Нью-Йорк) и Leducq Foundation (Париж, Франция). Исследование Women's Health Study выполнено при поддержке National Heart, Lung, and Blood Institute и National Cancer Institute (Бетесда, штат Мэриленд)» [208]. (Прогнозирование; Разработка.)

«Анализ данных частично выполнен при поддержке Clinical and Translational Service Center при Weill Cornell Medical College. Спонсоры не влияли на планирование нашего анализа, его интерпретацию и решение о направлении рукописи для публикации» [380]. (Диагностика; Разработка; Проверка.)

Пояснение

Предсказательные исследования, в том числе проспективные, как правило, получают финансирование в незначительном объёме или не получают его вовсе, что, как предполагается, способствует появлению большого количества исследований низкого качества. Многие из них проводятся без какой-либо экспертной оценки на этапе планирования, когда обычно запрашивается финансирование [472].

Авторы должны раскрывать все источники финансирования, полученного для исследования, и указывать роль спонсора в планировании, проведении, анализе и представлении результатов исследования. Если спонсоры не участвовали в этом, об этом также необходимо сообщить. Точно так же если исследование не получило внешнего финансирования, авторы должны чётко заявить об этом. Для моделей, включённых в клинические руководства, важно показать потенциальные финансовые и другие конфликты интересов всех участников разработки таких руководств, а не только тех, кто разрабатывал предсказательные модели [316, 483, 484].

ЗАКЛЮЧЕНИЕ

Исследования, посвящённые предсказательным моделям, многочисленны, а количество публикаций, описывающих разработку, проверку, обновление или расширение предсказательных моделей, не уменьшается. Руководство TRIPOD призвано предоставить полезные рекомендации по составлению отчётов об исследованиях по разработке или проверке (без обновления или с ним) одной или более предсказательных моделей для диагностических либо прогностических целей. Только при полной и прозрачной отчётности можно выявить сильные и слабые стороны исследования, что облегчит его интерпретацию и сделает его пригодным для использования [485–487]. Полная отчётность лежит в основе будущих исследований предсказательных моделей, в частности, позволяя исследователям проверять и сравнивать существующие модели. Полная отчётность также может способствовать принятию и внедрению проверенных предсказательных моделей для использования в повседневной практике. Руководство TRIPOD будет полезным для рецензентов и редакторов журналов при оценке статей об исследованиях, посвящённых предсказательным моделям. TRIPOD также может помочь при планировании, проведении и анализе исследований предсказательных моделей.

TRIPOD разработаны междисциплинарной группой из 24 экспертов, включая тех, кто принимал участие в разработке публикационных стандартов CONSORT [96], STROBE [97, 99], PRISMA [488], REMARK [98], GRIPS [101], STREGA [489], STARD [100], ARRIVE [490], CARE [491]. Используя этот коллективный опыт разработки руководств на основе консенсуса с экспертным знанием предмета, мы придерживались рекомендаций по разработке публикационных стандартов [113]. Мы обосновали и подробно обсудили каждый пункт контрольного перечня и привели наглядные примеры хорошей отчётности. По возможности мы ссылались на соответствующие эмпирические данные из обзоров публикаций. Кроме того, мы включили несколько блоков для дополнительного обсуждения основных вопросов разработки и проверки предсказательных моделей.

Некоторые могут возразить, что TRIPOD увеличит нагрузку на авторов, рецензентов и журналы. Мы же считаем, что использование TRIPOD, вероятно, сократит время рецензирования, уменьшит количество запросов на исправления и поможет обеспечить объективный процесс рецензирования [108]. Пункты, включённые в контрольный перечень, отражают многочисленные дискуссии, направленные на достижение консенсуса в отношении минимального объёма информации, которую необходимо представить, чтобы обеспечить информированную оценку качества исследования, рисков систематической ошибки и клинической значимости, а также сделать возможным использование результатов [532].

Существует ошибочное мнение о том, что публикационные руководства ограничивают творческий подход исследователей. TRIPOD, как и другие публикационные руководства, не содержит указаний, как проводить анализ, а скорее, описывает, как следует представлять его результаты.

Наконец, руководство TRIPOD следует рассматривать как развивающийся документ, требующий постоянной оценки и, если необходимо, уточнений, поскольку методология исследований предсказательных моделей продолжает развиваться. На веб-сайте TRIPOD (www.tripod-statement.org) будет форум для обсуждения, предложений по совершенствованию контрольного перечня и настоящего документа с пояснениями и уточнениями, а также дополнительной информацией, касающейся исследований предсказательных моделей. Мы также планируем поощрять перевод контрольного списка на другие языки и планируем размещать их на нашем веб-сайте. Объявления и информация, касающиеся TRIPOD, будут доступны на странице руководства в Twitter (@TRIPODStatement). TRIPOD также будет связан с библиотекой EQUATOR Network и будет продвигаться ею с целью повышения качества и прозрачности отчётности об исследованиях в области здравоохранения (www.equator-network.org).

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

Благодарности. При участии Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht (Утрехт, Нидерланды); Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford (Оксфорд, Великобритания); Stanford Prevention Research Center, School of Medicine, School of Humanities and Sciences, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University (Стэнфорд, Калифорния); Screening and Test Evaluation Program (STEP), School of Public Health, Sydney Medical School, University of Sydney (Сидней, Австралия); Erasmus MC-University Medical Center Rotterdam (Роттердам, Нидерланды); Memorial Sloan Kettering Cancer Center (Нью-Йорк, штат Нью-Йорк, США); University of North Carolina at Chapel Hill (Чапел-Хилл, штат Северная Каролина, США).

Разглашение сведений: с информацией можно ознакомиться на сайте www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M14-0698.

Грантовая поддержка. Разработка контрольного перечня и рекомендаций выполнена без прямой спонсорской поддержки. Заседание в июне 2011 г. проведено при частичной поддержке National Institute for Health Research Senior Investigator Award во главе с доктором Altman, а также грантов со стороны Cancer Research UK (C5529) и Netherlands Organization for Scientific Research (ZONMW 918.10.615 и 91208004). Доктора Collins и Altman получили частичную финансовую поддержку со стороны Medical Research Council (G1100513). Д-р Altman является членом Medical Research Council Prognosis Research Strategy (PROGRESS) Partnership (G0902393/99558).

Запросы на однократное переиздание: Karel G.M. Moons, PhD, Julius Centre for Health Sciences and Primary Care, UMC Utrecht. Почтовый адрес: PO Box 85500, 3508 GA Utrecht, the Netherlands. E-mail: K.G.M.Moons@umcutrecht.nl

Вклад авторов: K.G.M. Moons, D.G. Altman , J.B. Reitsma, P. Macaskill, G.S. Collins — концепция и планирование. K.G.M. Moons, D.G. Altman , J.B. Reitsma, J.P.A. Loannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins — анализ и интерпретация данных. K.G.M. Moons, D.G. Altman , J.B. Reitsma, G.S. Collins — подготовка черновика статьи. K.G.M. Moons, D.G. Altman , J.B. Reitsma, J.P.A. Loannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins — критическая доработка статьи с внесением важного интеллектуального содержания. K.G.M. Moons, D.G. Altman , J.B. Reitsma, J.P.A. Loannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins — окончательное утверждение статьи. K.G.M. Moons, D.G. Altman , J.B. Reitsma, G.S. Collins — поиск материалов исследования или пациентов. K.G.M. Moons, D.G. Altman, J.B. Reitsma, P. Macskill, E.W. Steyerberg, A.J. Vickers, G.S. Collins — статистическая экспертиза. K.G.M. Moons, D.G. Altman, G.S. Collins — получение финансирования. K.G.M. Moons, G.S. Collins — административная, техническая или логистическая поддержка. K.G.M. Moons, D.G. Altman, G.S. Collins — сбор и объединение данных.

Члены группы TRIPOD

Gary Collins (University of Oxford, Оксфорд, Великобритания); Douglas Altman (University of Oxford, Оксфорд, Великобритания); Karel Moons (University Medical Center Utrecht, Утрехт, Нидерланды); Johannes Reitsma (University Medical Center Utrecht, Утрехт, Нидерланды); Virginia Barbour *(PLoS Medicine,* Великобритания, Австралия); Nancy Cook (Division of Preventive Medicine, Brigham & Women's Hospital, Бостон, штат Массачусетс, США); Joris de Groot (University Medical Center Utrecht, Утрехт, Нидерланды); Trish Groves *(BMJ,* Лондон, Великобритания); Frank Harrell Jr. (Vanderbilt University, Нашвилл, штат Теннесси, США); Harry Hemingway (University College London, Лондон, Великобритания); John Ioannidis (Stanford University, Стэнфорд, штат Калифорния, США); Michael W. Kattan (Cleveland Clinic, Кливленд, штат Огайо, США); André

Knottnerus (Maastricht University, Маастрихт, Нидерланды, и *Journal of Clinical Epidemiology);* Petra Macaskill (University of Sydney, Сидней, Австралия); Susan Mallett (University of Oxford, Оксфорд, Великобритания); Cynthia Mulrow *(Annals of Internal Medicine,* American College of Physicians, Филадельфия, штат Пенсильвания, США); David Ransohoff (University of North Carolina at Chapel Hill, Чапел-Хилл, штат Северная Каролина, США); Richard Riley (University of Birmingham, Бирмингем, Великобритания); Peter Rothwell (University of Oxford, Оксфорд, Великобритания); Patrick Royston (Medical Research Council Clinical Trials Unit at University College London, Лондон, Великобритания); Willi Sauerbrei (University of Freiburg, Фрайбург, Германия); Ewout Steyerberg (University Medical Center Rotterdam, Роттердам, Нидерланды); Ian Stiell (University of Ottawa, Оттава, провинция Онтарио, Канада); Andrew Vickers (Memorial Sloan Kettering Cancer Center, Нью-Йорк, США).

СПИСОК ЛИТЕРАТУРЫ

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

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.

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.

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.

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

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.

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.

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.

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

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.

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

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

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.

15. Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Möhlig 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.

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.

17. D'Agostino RB, 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.

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.

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

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.

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.

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

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.

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.

25. Steyerberg EW, Eijkemans MJ, Harrell FE, 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.

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

27. Ioannidis JPA, Khoury MJ. Improving validation practices in "omics" research. Science. 2011;334:1230-2.

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

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.
30. Taylor JM, Ankesrt DP, Andridge RR. Validation of biomarker-based risk prediction models. Clin Cancer Res. 2008;14:5977-83.
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.
32. Steyerberg EW, Harrell FE, 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.
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.
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.
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.
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.
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.
40. Lackland DT, Elkind MS, D'Agostino R, Dhamoon MS, Goff DC, 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.
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.
42. Shariat SF, Karakiewicz PI, Margulis V, Kattan MW. Inventory of prostate cancer predictive tools. Curr Opin Urol. 2008;18:279-96.
43. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest. 2009;27:235-43.
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.
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.
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.
47. Collins GS, Moons KG. Comparing risk prediction models. BMJ. 2012;344:e3186.
48. Siontis GC, Tzoulaki I, Siontis KC, Ioannidis JP. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012;344:e3318.
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.
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.
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.
52. Groves T, Godlee F. Open science and reproducible research. BMJ. 2012;344:e4383.
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.
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.
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.
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.
57. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med. 1993;118:201-10.
58. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA. 1997;277:488-94.
59. Steurer J, Haller C, Häuselmann 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.
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.
61. Hayden JA, Côté P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med. 2006;144:427-37.

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.
63. Mushkudiani NA, Hukkelhoven CW, Hernández 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.
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.
65. Siontis GC, Tzoulaki I, Ioannidis JP. Predicting death: an empirical evaluation of predictive tools for mortality. Arch Intern Med. 2011;171:1721-6.
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.
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.
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.
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.
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.
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.
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.
73. Echouffo-Tcheugui JB, Kengne AP. Risk models to predict chronic kidney disease and its progression: a systematic review. PLoS Med. 2012;9:e1001344.
74. Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP. Risk models to predict hypertension: a systematic review. PLoS One. 2013;8:e67370.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
89. Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14:3.
90. Huen SC, Parikh CR. Predicting acute kidney injury after cardiac surgery: a systematic review. Ann Thorac Surg. 2012;93:337-41.
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.
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.
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.
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.
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.
96. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.
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.

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.
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.
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.
101. Janssens AC, Ioannidis JP, vanDuijn 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.
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.
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.
104. Labarère 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.
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.
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.
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.
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.
109. Campbell MK, Elbourne DR, Altman DG; CONSORT Group. CONSORT statement: extension to cluster randomised trials. BMJ. 2004;328:702-8.
110. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet. 1974;2:81-4.
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.
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.
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.
116. Dehing-Oberije C, Yu S, DeRuysscher D, Meersschout S, VanBeek 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.
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.
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.
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.
120. D'Agostino RB, 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.
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.
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.
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.
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.
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.
126. Ingui BJ, Rogers MA. Searching for clinical prediction rules in . J Am Med Inform Assoc. 2001;8:391-7.
127. Wong SS, Wilczynski NL, Haynes RB, Ramkissoonsingh R; Hedges Team. Developing optimal search strategies for detecting sound clinical prediction studies in . AMIA Annu Symp Proc. 2003:728-32.
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.
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 to establish a Web-based register. J Clin Epidemiol. 2011;64:848-60.
130. Rietveld RP, terRiet 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.
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.
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.
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.
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.
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.
136. Büller 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.
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.
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.
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.
140. Conroy RM, Pyörälä 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, 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.
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.
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.
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.
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.
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.
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.
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.
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.
153. Knottnerus JA, Dinant GJ. Medicine based evidence, a prerequisite for evidence based medicine. BMJ. 1997;315:1109-10.
154. Knottnerus JA, vanWeel C, Muris JW. Evaluation of diagnostic procedures. BMJ. 2002;324:477-80.
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.
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.
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.
158. Lumbreras B, Parker LA, Porta M, Pollán M, Ioannidis JP, Hernández-Aguado I. Overinterpretation of clinical applicability in molecular diagnostic research. Clin Chem. 2009;55:786-94.
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.
160. Greving JP, Wermer MJ, Brown RD, 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.
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.
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.
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.

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.
165. Debray TP, Moons KG, Abo-Zaid GM, Koffijberg H, Riley RD. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? PLoS One. 2013;8:e60650.
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.
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.
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.
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.
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.
171. van Klaveren D, Steyerberg EW, Vergouwe Y. Interpretation of concordance measures for clustered data. Stat Med. 2014;33:714-6.
172. Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L. Derivation and assessment of risk prediction models using case-cohort data. BMC Med Res Methodol. 2013;13:113.
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.
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.
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.
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.
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.
178. Ronga A, Vaucher P, Haasenritter J, Donner-Banzhoff N, Bösner 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.
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.
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.
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.
182. Counsell C, Dennis M. Systematic review of prognostic models in patients with acute stroke. Cerebrovasc Dis. 2001;12:159-70.
183. Knottnerus JA. Between iatrotropic stimulus and interiatric referral: the domain of primary care research. J Clin Epidemiol. 2002;55:1201-6.
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.
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.
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.
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.
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.
189. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol. 1992;45:1143-54.
190. Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and diseases. Med Decis Making. 1987;7:139-48.
191. Eberhart LH, Morin AM, Guber D, Kretz FJ, Schäuffelen 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.
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].
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 Sézary syndrome. Br J Dermatol. 2005;153:118-24.
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.
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.
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.
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.
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.

199. Moons KG, Donders RA, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59:1092-101.
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.
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.
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.
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.
204. Simon R, Altman D, G. Statistical aspects of prognostic factor studies in oncology. Br J Cancer. 1994;69:979-85.
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.
206. Schuit E, Groenwold RH, Harrell FE, 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.
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.
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.
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.
210. Massing MW, Simpson RJ, 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.
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.
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.
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.
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.
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.
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.
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.
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.
220. Rutjes AW, Reitsma JB, DiNisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006;174:469-76.
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.
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.
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.
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.
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.
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.
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.
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.
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.
230. Beyersmann J, Wolkewitz M, Schumacher M. The impact of time-dependent bias in proportional hazards modelling. Stat Med. 2008;27:6439-54.
231. van Walraven C, Davis D, Forster AJ, Wells GA. Time-dependent bias was common in survival analyses published in leading clinical journals. J Clin Epidemiol. 2004;57:672-82.
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.

234. Scheike TH. Time-varying effects in survival analysis.. In: Rao CR, eds. 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.
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.
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.
238. Begg CB. Bias in the assessment of diagnostic tests. Stat Med. 1987;6:411-23.
239. Elmore JG, Wells CK, Howard DH, Feinstein AR. The impact of clinical history on mammographic interpretations. JAMA. 1997;277:49-52.
240. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. JAMA. 2004;292:1602-9.
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.
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.
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.
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.
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.
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.
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.
248. Feinstein AR. Multivariable Analysis. New Haven, CT: Yale University Press; 1996.
249. Schumacher M, Holländer 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.
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.
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.
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.
255. Audigé 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.
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.
257. Thompson DO, Hurtado TR, Liao MM, Byyny RL, Gravitz C, Haukoos JS. Validation of the Simplified Motor Score in the out-of-hospital setting for the prediction of outcomes after traumatic brain injury. Ann Emerg Med. 2011;58:417-25.
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.
259. Mackinnon A. The use and reporting of multiple imputation in medical research—a review. J Intern Med. 2010;268:586-93.
260. Hussain A, Dunn KW. Predicting length of stay in thermal burns: a systematic review of prognostic factors. Burns. 2013;39:1331-40.
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.
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.
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.
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.
265. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25:127-41.
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.
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.
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.

270. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008;19:640-8.
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.
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.
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.
280. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. Am Heart J. 1991;121:293-8.
281. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21:2175-97.
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.
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.
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.
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.
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.
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.
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.
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.
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 high-dimensional data. J Biopharm Stat. 2011;21:1206-31.
293. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med. 1984;3:143-52.
294. van Houwelingen JC, LeCessie S. Predictive value of statistical models. Stat Med. 1990;9:1303-25.
295. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005;21:3301-7.
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.
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.
299. Castaldi PJ, Dahabreh IJ, Ioannidis JP. An empirical assessment of validation practices for molecular classifiers. Brief Bioinform. 2011;12:189-202.
300. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91.
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.
303. Ambler G, Seaman S, Omar RZ. An evaluation of penalised survival methods for developing prognostic models with rare events. Stat Med. 2012;31:1150-61.
304. Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. Prognostic indices for older adults: a systematic review. JAMA. 2012;307:182-92.
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.
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.
307. Meffert PJ, Baumeister SE, Lerch MM, Mayerle J, Kratzer W, Völzke H. Development, external validation, and comparative assessment of a new diagnostic score for hepatic steatosis. Am J Gastroenterol. 2014;109:1404-14.
308. Collins GS, Altman DG. Identifying patients with undetected colorectal cancer: an independent validation of QCancer (Colorectal). Br J Cancer. 2012;107:260-5.

309. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33.
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.
311. Zivanovic O, Jacks LM, Iasonos A, Leitao MM, 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.
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.
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.
314. Collins GS, Michaëlsson K. Fracture risk assessment: state of the art, methodologically unsound, or poorly reported? Curr Osteoporos Rep. 2012;10:199-207.
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.
316. Järvinen TL, Jokihaara J, Guy P, Alonso-Coello P, Collins GS, Michaëlsson K, et al. Conflicts at the heart of the FRAX tool. CMAJ. 2014;186:165-7.
317. Balmaña 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.
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.
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.
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.
321. Akazawa K. Measures of explained variation for a regression model used in survival analysis. J Med Syst. 1997;21:229-38.
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.
323. Heller G. A measure of explained risk in the proportional hazards model. Biostatistics. 2012;13:315-25.
324. Korn EL, Simon R. Measures of explained variation for survival data. Stat Med. 1990;9:487-503.
325. Mittlböck M, Schemper M. Explained variation for logistic regression. Stat Med. 1996;15:1987-97.
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.
328. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. Biometrics. 2000;56:249-55.
329. Schemper M, Stare J. Explained variation in survival analysis. Stat Med. 1996;15:1999-2012.
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.
331. Rufibach K. Use of Brier score to assess binary predictions. J Clin Epidemiol. 2010;63:938-9.
332. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biom J. 2008;50:457-79.
333. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med. 2004;23:723-48.
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.
335. Demler OV, Pencina MJ, D'Agostino RB. Misuse of DeLong test to compare AUCs for nested models. Stat Med. 2012;31:2577-87.
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.
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.
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.
339. Pencina MJ, D'Agostino RB, D'Agostino RB, 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.
340. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115:928-35.
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.
342. Cook NR. Assessing the incremental role of novel and emerging risk factors. Curr Cardiovasc Risk Rep. 2010;4:112-9.
343. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol. 2011;11:13.
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.
345. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. Biom J. 2011;53:237-58.
346. Cook NR. Clinically relevant measures of fit? A note of caution. Am J Epidemiol. 2012;176:488-91.
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.
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.
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.
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.
351. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. Am J Epidemiol. 2011;173:1327-35.
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.
353. Mühlenbruch 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.
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. UW Biostatistics 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.
356. Hilden J. Commentary: On NRI, IDI, and "good-looking" statistics with nothing underneath. Epidemiology. 2014;25:265-7.
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.
358. Al-Radi OO, Harrell FE, 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.
359. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med. 2012;157:294-5.
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].
361. Vickers AJ. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. Am Stat. 2008;62:314-20.
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.
363. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74.
364. Baker SG. Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst. 2009;101:1538-42.
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.
366. Baker SG, Kramer BS. Evaluating a new marker for risk prediction: decision analysis to the rescue. Discov Med. 2012;14:181-8.
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.
368. Held U, Bové DS, Steurer J, Held L. Validating and updating a risk model for pneumonia—a case study. BMC Med Res Methodol. 2012;12:99.
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.
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.
371. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009;374:86-9.
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.
373. van Houwelingen HC. Validation. calibration, revision and combination of prognostic survival models. Stat Med. 2000;19:3401-15.
374. Manola J, Royston P, Elson P, McCormack JB, Mazumdar M, Négrier 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.
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.
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 IgM-related disorders. Leukemia. 2004;18:1512-7.
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.
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.
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.
380. Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA, et al. Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med. 2009;151:775-83.
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.
382. Starmans R, Muris JW, Fijten GH, Schouten HJ, Pop P, Knottnerus JA. The diagnostic value of scoring models for organic and non-organic gastrointestinal disease, including the irritable-bowel syndrome. Med Decis Making. 1994;14:208-16.
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.

384. Harrison DA, Rowan KM. Outcome prediction in critical care: the ICNARC model. Curr Opin Crit Care. 2008;14:506-12.
385. Kanaya AM, WasselFyr 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.
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.
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.
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.
389. Geersing GJ, Erkens PM, Lucassen WA, Büller 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.
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.
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.
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.
393. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials. 1996;17:343-6.
394. Echouffo-Tcheugui JB, Woodward M, Kengne AP. Predicting a post-thrombolysis intracerebral hemorrhage: a systematic review. J Thromb Haemost. 2013;11:862-71.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
413. Cole TJ. Scaling and rounding regression-coefficients to integers. Appl Stat. 1993;42:261-8.
414. Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: the Framingham study risk score functions. Stat Med. 2004;23:1631-60.
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.

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.
417. Royston P, Altman DG. Visualizing and assessing discrimination in the logistic regression model. Stat Med. 2010;29:2508-20.
418. Taş 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.
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.
420. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11-21.
421. Pepe MS, Janes H. Reporting standards are needed for evaluations of risk reclassification. Int J Epidemiol. 2011;40:1106-8.
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.
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.
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.
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.
426. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. J Clin Epidemiol. 2007;60:324-9.
427. Horton R. The hidden research paper. JAMA. 2002;287:2775-8.
428. Docherty M, Smith R. The case for structuring the discussion of scientific papers. BMJ. 1999;318:1224-5.
429. Ioannidis JP. Research needs grants, funding and money—missing something? Eur J Clin Invest. 2012;42:349-51.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
441. Ioannidis JP. More than a billion people taking statins? Potential implications of the new cardiovascular guidelines. JAMA. 2014;311:463-4.
442. Ioannidis JP, Tzoulaki I. What makes a good predictor? The evidence applied to coronary artery calcium score. JAMA. 2010;303:1646-7.
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.
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.
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.
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.
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.
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.
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.
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.

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.
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.
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.
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.
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.
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.
458. Pita-Fernández S, Pértega-Diaz S, Valdés-Cañedo F, Seijo-Bestilleiro R, Seoane-Pillado T, Fernández-Rivera C, et al. Incidence of cardiovascular events after kidney transplantation and cardiovascular risk scores: study protocol. BMC Cardiovasc Disord. 2011;11:2.
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.
460. Siebeling L, terRiet 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.
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.
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.
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.
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 randomised controlled trial using a web-based risk formatter (protocol). BMC Med Inform Decis Mak. 2010;10:41.
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.
466. Peng RD. Reproducible research and *Biostatistics.* Biostatistics. 2009;10:405-8.
467. Keiding N. Reproducible research and the substantive context. Biostatistics. 2010;11:376-8.
468. Vickers AJ. Whose data set is it anyway? Sharing raw data from randomized trials. Trials. 2006;7:15.
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.
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.
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.
472. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. BMJ. 2009;339:b4184.
473. Groves T. BMJ policy on data sharing. BMJ. 2010;340:c564.
474. Marchionni L, Afsari B, Geman D, Leek JT. A simple and reproducible breast cancer prognostic test. BMC Genomics. 2013;14:336.
475. Loder E, Groves T, Macauley D. Registration of observational studies. BMJ. 2010;340:c950.
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.
477. Should protocols for observational studies be registered? Lancet. 2010;375:348.
478. Altman DG. The time has come to register diagnostic and prognostic research. Clin Chem. 2014;60:580-2.
479. The registration of observational studies—when metaphors go bad. Epidemiology. 2010;21:607-9.
480. Sørensen HT, Rothman KJ. The prognosis of research. BMJ. 2010;340:c703.
481. Vandenbroucke JP. Registering observational research: second thoughts. Lancet. 2010;375:982-3.
482. Williams RJ, Tse T, Harlan WR, Zarin DA. Registration of observational studies: Is it time? CMAJ. 2010;182:1638-42.
483. Lenzer J. Majority of panelists on controversial new cholesterol guideline have current or recent ties to drug manufacturers. BMJ. 2013;347:f6989.
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.
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.
486. Simera I, Kirtley S, Altman DG. Reporting clinical research: guidance to encourage accurate and transparent research reporting. Maturitas. 2012;72:84-7.
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.
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.
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.
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.
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.
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.
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.
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.
497. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8:3-15.
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.
499. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18:681-94.
500. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Stat Med. 2008;27:3227-46.
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.
503. Vickers AJ, Lilja H. Cutpoints in clinical chemistry: time for fundamental reassessment. Clin Chem. 2009;55:15-7.
504. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol. 2012;12:21.
505. Dawson NV, Weiss R. Dichotomizing continuous variables in statistical analysis: a practice to avoid. Med Decis Making. 2012;32:225-6.
506. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat. 1994;43:429-67.
507. Harrell FE, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst. 1988;80:1198-202.
508. Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. Bioinformatics. 2007;23:1768-74.
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.
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.
511. Boulesteix AL. Validation in bioinformatics and molecular medicine. Brief Bioinform. 2011;12:187-8.
512. Jelizarow M, Guillemot V, Tenenhaus A, Strimmer K, Boulesteix AL. Over-optimism in bioinformatics: an illustration. Bioinformatics. 2010;26:1990-8.
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.
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.
515. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Methods Med Res. 2014 Apr 7 [Epub ahead of print]..
516. Vach W. Calibration of clinical prediction rules does not just assess bias. J Clin Epidemiol. 2013;66:1296-301.
517. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic-regression models. Stat Med. 1991;10:1213-26.
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.
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.
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.
523. Marcin JP, Romano PS. Size matters to a model's fit. Crit Care Med. 2007;35:2212-3.
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.
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.
526. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. 2008;54:17-23.

527. Pencina MJ, D'Agostino RB, Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. Stat Med. 2012;31:1543-53.
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.
529. Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics. 2014;15:526-39.
530. Pencina MJ, D'Agostino RB, 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.
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.
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.

ОБ АВТОРАХ:

Karel G.M. Moons, PhD; email: K.G.M.Moons@umcutrecht.nl; ORCID: https://orcid.org/0000-0003-2118-004X;
address: Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands

Douglas G. Altman, DSc; ORCID: https://orcid.org/0000-0002-7183-4083

Johannes B. Reitsma, MD, PhD; ORCID: https://orcid.org/0000-0003-4026-4345

John P.A. Loannidis, MD, DSc; email: jioannid@stanford.edu; ORCID: https://orcid.org/0000-0003-3118-6859

Petra Macaskill, PhD; email: petra.macaskill@sydney.edu.au; ORCID: https://orcid.org/0000-0001-5879-6193

Ewout W. Steyerberg, PhD; email: e.w.steyerberg@lumc.nl; ORCID: https://orcid.org/0000-0002-7787-0122

Andrew J. Vickers, PhD; email: vickersa@mskcc.org; ORCID: https://orcid.org/0000-0003-1525-6503

David F. Ransohoff, MD; email: ransohof@med.unc.edu; ORCID: https://orcid.org/0000-0002-2200-039X

Gary S. Collins, PhD; email: gary.collins@csm.ox.ac.uk; ORCID: https://orcid.org/0000-0002-2772-2316