ОБЗОРЫ

REVIEWS

Прогностические модели в медицине

А.С. Лучинин

ФГБУН «Кировский НИИ гематологии и переливания крови ФМБА», ул. Красноармейская, д. 72, Киров, Российская Федерация, 610027

РЕФЕРАТ

Медицинские прогностические (предиктивные) модели (МПМ) имеют важное значение в современном здравоохранении. Они определяют риски для здоровья и возникновения заболеваний. Целью их создания является улучшение результатов диагностики и лечения. Все МПМ можно разделить на две категории. Диагностические медицинские модели (ДММ) помогают рассчитать индивидуальный риск присутствия заболевания, в то время как прогностические медицинские модели (ПММ) — риск возникновения болезни или его осложнения в будущем. В обзоре обсуждаются характеристики ДММ и ПММ, условия их разработки, критерии применения в медицине, в частности в гематологии, а также проблемы, возникающие на этапе их создания и проверки качества.

Ключевые слова: прогностическая модель, искусственный интеллект.

Получено: 13 сентября 2022 г.
Принято в печать: 7 декабря 2022 г.

Для переписки: Александр Сергеевич Лучинин, канд. мед. наук, ул. Красноармейская, д. 72, Киров, Российская Федерация, 610027; тел.: +7(919)506-87-86; e-mail: glivec@mail.ru

Для цитирования: Лучинин А.С. Прогностические модели в медицине. Клиническая онкогематология. 2023;16(1):27–36.

DOI: 10.21320/2500-2139-2023-16-1-27-36

Prognostic Models in Medicine

AS Luchinin

Kirov Research Institute of Hematology and Transfusiology, 72 Krasnoarmeiskaya ul., Kirov, Russian Federation, 610027

ABSTRACT

Medical prognostic (prediction) models (MPM) are essential in modern healthcare. They determine health and disease risks and are created to improve diagnosis and treatment outcomes. All MPMs fall into two categories. Diagnostic medical models (DMM) aim at assessing individual risk for a disease present, whereas predictive medical models (PMM) evaluate the risk for development of a disease and its complications in future. This review discusses DMM and PMM characteristics, conditions for their elaboration, criteria for medical application, also in hematology, as well as challenges of their creation and quality check.

Keywords: prognostic model, artificial intelligence.

Received: September 13, 2022
Accepted: December 7, 2022

For correspondence: Aleksander Sergeevich Luchinin, MD, PhD, 72 Krasnoarmeiskaya ul., Kirov, Russian Federation, 610027; Tel.: +7(919)506-87-86; e-mail: glivec@mail.ru

For citation: Luchinin AS. Prognostic Models in Medicine. Clinical oncohematology. 2023;16(1):27–36. (In Russ).

DOI: 10.21320/2500-2139-2023-16-1-27-36

ВВЕДЕНИЕ

Ежегодно в медицинской литературе появляются сотни новых прогностических статистических моделей. Только за 9 мес. пандемии опубликовано 232 новые системы прогноза, связанные с COVID-19 [1]. Медицинские прогностические (предиктивные) модели (МПМ) относят к одной из двух основных категорий: диагностические медицинские модели (ДММ) для оценки состояния здоровья человека в текущий момент времени, например риска заболевания, и прогностические медицинские модели (ПММ), которые помогают рассчитать вероятность развития того или иного исхода в течение определенного периода времени [2].

Примерами ДММ, которые в гематологии распространены гораздо реже, чем ПММ, служат модель дифференциальной диагностики эссенциальной тромбоцитемии и префибротического миелофиброза, построенная на базе логистической регрессии, и система диагностики заболеваний крови, основанная на резуль-

татах лабораторных анализов [3, 4]. Хорошо известными и используемыми ПММ являются такие, как R-IPI и IPI (для оценки прогноза течения диффузной В-крупноклеточной лимфомы), FLIPI (для фолликулярной лимфомы), R-ISS (для множественной миеломы) и многие другие [5–7]. МПМ разного качества существуют практически для всех известных заболеваний человека и их осложнений.

В последние годы создается большое количество прогностических моделей в гематологии с использованием технологий машинного обучения (МО) для диагностики заболеваний, прогнозирования осложнений и результатов лечения [8]. Алгоритмы МО являются перспективными с технической точки зрения, но далеко не все они доходят до внедрения в реальную клиническую практику.

Результат МПМ — вероятность (риск) исхода, который может быть выражен в виде числа от 0 до 1 (или от 0 до 100 %) или условной группы риска (например, низкий, промежуточный или высокий риск). За исход принимается событие, например смерть пациента или осложнение заболевания, либо время до их наступления (общая или бессобытийная выживаемость). МПМ позволяют рассчитать отношение рисков, или отношение шансов, или относительный риск. Эти параметры не указывают напрямую на вероятность исхода, хотя она может быть при этом рассчитана. Все относительные показатели отражают изменение риска исхода между сравниваемыми группами в условиях наличия либо отсутствия тех или иных факторов.

В ДММ результатом прогнозирования является текущее состояние здоровья пациента в данный момент времени. Создание ДММ — анализ, при котором данные пациентов размечаются по наличию или отсутствию у них целевого диагноза. При этом ведется учет предикторов в момент или с небольшим промежутком времени до либо после постановки диагноза. Примером может служить модель предварительной диагностики рака легкого по результатам компьютерной томографии с применением технологий искусственного интеллекта, когда диагноз впоследствии подтверждается или опровергается по данным биопсии [9].

При создании ПММ внимание уделяется будущему результату для здоровья, который произойдет после прогнозирования с использованием доступных на тот момент предикторов. Применяется такое понятие, как горизонт прогнозирования — насколько далеко во времени модель стремится предсказать изучаемый исход. При создании ПММ период наблюдения за пациентами должен соответствовать данному горизонту, который может варьировать от нескольких дней до нескольких лет. Пациенты, у которых изучаемый результат не зарегистрировали до конца горизонта прогнозирования, т. к. они либо еще его не достигли, либо выбыли из-под наблюдения, подвергаются цензурированию, что, например, учитывается в моделях выживаемости [2].

Основные термины и их интерпретация

ФАКТОР РИСКА — любая характеристика, которая может менять вероятность результата (исхода) для пациента.

ВЕРОЯТНОСТЬ ИЛИ РИСК — количественная оценка возможности наступления определенного события, например болезни или смерти, выраженная в виде числа от 0 до 1 (или от 0 до 100 %).

ОТНОСИТЕЛЬНЫЙ РИСК — отношение риска наступления определенного события у пациентов, подвергшихся воздействию фактора риска, по отношению к пациентам контрольной группы, не подвергавшихся влиянию этого фактора.

ШАНСЫ — отношение числа наблюдений с достигнутым исходом к числу наблюдений без него в рамках горизонта прогнозирования.

ОТНОШЕНИЕ ШАНСОВ (ОШ) — шансы того, что исход произойдет при определенном воздействии, по сравнению с шансами того, что исход произойдет в его отсутствие. При равных шансах значение их отношения равно 1, а вероятность исхода — 0,5 (50 %) [10].

ОТНОШЕНИЕ РИСКОВ (ОТНОШЕНИЕ ОПАСНОСТЕЙ ИЛИ УГРОЗ) — отношение рисков исхода в двух группах пациентов, которые различаются между собой факторами риска, где под риском понимается частота возникновения интересующего события, например смерти, за определенный промежуток времени. Показатель отражает разницу между риском исхода для пациента в одной группе с риском исхода для пациента в другой группе в один и тот же следующий момент времени на протяжении всего горизонта прогнозирования при условии, что пациенты еще не достигли данного результата [11].

СОЗДАНИЕ ПРОГНОСТИЧЕСКОЙ МОДЕЛИ

Разработка МПМ проходит в несколько этапов анализа данных и оценки качества полученных результатов. Они включают в себя описание характеристик будущей модели, таких как прогнозируемый результат, целевая популяция и критерии использования, подготовку соответствующего набора данных, выбор алгоритма моделирования, например логистической регрессии, поиск кандидатов-предикторов и др. [12].

В ДММ и ПММ основное внимание уделяется не эффектам отдельных предикторов, а эффективности модели в целом. Иногда цель исследования — изучение возможности нового предиктора улучшить существующую модель. Эффективность модели для клинического прогнозирования оценивается через параметры дискриминативности и корректируется посредством калибровки. Дискриминативность — способность модели отличать пациента с исходом и без него [12]. Калибровка — коррекция предсказанных и наблюдаемых вероятностей для каждого из прогнозируемых классов данных. Калибровкой часто пренебрегают, но она имеет решающее значение, особенно при определении порогов рисков в принятии клинических решений [13].

Оценка эффективности МПМ на данных, которые применялись при ее разработке, приведет к ошибочно оптимистичным результатам [2]. Следовательно, во

время разработки модели требуется внутренняя проверка для оценки и корректировки модели с использованием валидационной выборки, перекрестной проверки (кросс-валидации) или бутстрэппинга. Внешняя проверка на данных из другой совокупности, которые не использовались для разработки, должна выполняться до практического применения модели. Важно, чтобы набор данных для разработки имел достаточно большой размер, соответствующий сложности модели. Данные должны быть высокого качества с беспристрастной выборкой участников, например последовательной, а процесс моделирования должен быть избавлен от предвзятости и рисков систематического смещения в выгодную исследователю сторону [14]. Поскольку эффективность моделей будет различаться в зависимости от времени и места их применения, они требуют периодической повторной оценки (проверки во времени) и применения в различных медицинских учреждениях (географической проверки). По результатам внешней валидации может потребоваться перекалибровка или обновление модели.

Цель МПМ состоит в улучшении процесса принятия медицинских решений, при этом ее высокая эффективность в рамках научного эксперимента не является гарантией практического результата. Оценка способности тех или иных МПМ улучшать качество медицинской помощи для пациентов должна осуществляться в дополнительных исследованиях, которые до сих пор проводятся крайне редко [15].

Формулирование проблемы и проверка данных

Прежде чем приступить к созданию МПМ, исследователю необходимо решить несколько задач. Следует сформулировать четкую цель прогнозирования. Это может быть выявление прогностических факторов (предикторов). Для оценки значимости предикторов их эффекты обычно выражаются в относительной шкале, например в виде ОШ. Если целью является создание модели, то результат ее работы может быть представлен в виде вероятности в абсолютной шкале риска от 0 до 100 %. Неопределенность прогноза указывается в виде 95%-го доверительного интервала.

Обязательно следует провести анализ литературы по существу проблемы, часто требуется взаимодействие между статистиками и клиническими исследователями, имеющими экспертные знания в области практического применения будущей МПМ. Следует изучить, какие МПМ для решения поставленной задачи уже существуют и описаны в научной литературе, в чем их недостатки или преимущества, статус их практического применения к настоящему времени. Создание новой прогностической модели, не имеющей преимуществ перед существующими аналогами, нецелесообразно. При наличии моделей прогнозирования для того же исхода и целевой группы, которые уже используются в клинической практике, оптимальным решением является оценка возможностей их обновления, оптимизации и совершенствования, а также внешняя валидация.

Создание МПМ невозможно без качественных данных. Часто данные о пациентах, используемые в МПМ, исходно собираются для других целей в рамках иных исследований. В связи с этим важно учитывать их актуальность и репрезентативность на момент анализа. Прогноз выживаемости пациентов 20-летней давности не будет отражать истинный прогноз за последние несколько лет, т. к. характер лечения изменился, что называется «эффект влияния лечения», который часто игнорируется [12].

Качество информации характеризуется ее полнотой и надежностью измерения. Данные реальной клинической практики, связанные с клинической картиной заболевания, могут быть сознательно или несознательно искажены вследствие опечаток при внесении информации в медицинские информационные системы либо фальсифицированы в ходе рутинных процессов, связанных с копированием предыдущих медицинских записей. Многие наборы ретроспективных данных являются неполными в отношении значений некоторых потенциальных предикторов. По умолчанию пациенты с пропущенными значениями исключаются из статистического анализа (полный анализ случаев или доступный анализ случаев), но такой подход неэффективен, поскольку теряется доступная информация о других предикторах. Для решения проблемы существуют разные способы вменения пропущенных значений, в т. ч. с использованием методов МО. Вменение следует проводить осторожно, тщательно выбирая метод и оценивая результат, т. к. синтетические данные могут искажать истинное распределение. Тем не менее такой подход предпочтительнее полного анализа случаев [16]. Проспективные прогностические исследования, как правило, лишены указанных недостатков.

Исследователи, планирующие или разрабатывающие новые многопараметрические модели прогнозирования, должны учитывать все требования к размеру выборки для своего набора данных. Эти требования могут отличаться у моделей с разными типами результатов, например бинарными, и исходами во времени, такими как летальность и выживаемость. Главным расчетным показателем является EPP (events per predictor) — число событий на один предиктор. В частности, в моделях прогноза летальности или выживаемости событием является смерть пациента. Общий размер выборки вычисляется из величины параметра EPP с учетом известной или прогнозируемой частоты данного события. Низкий показатель EPP приводит к серьезным проблемам моделирования, таким как смещение коэффициентов регрессии как в положительную, так и отрицательную сторону, к завышению их дисперсии, некорректным доверительным интервалам и парадоксам, например инверсии коэффициентов регрессии (парадокс Симпсона) [17].

Существуют различные способы расчета минимального размера выборки для моделей прогноза. Правило большого пальца (the rule of thumb) — эмпирический подход, позволяющий получить приблизительный результат. Часто используемое эмпирическое правило для размера выборки состоит в том, чтобы обеспечить не менее 10 событий на один предиктор (переменную), который включается в окончательную модель. Если предиктор имеет 3 категории (пара-

метра) и более, то число событий должно рассчитываться на количество этих категорий минус 1, поэтому следует использовать показатель EPP, а не число событий на одну переменную (events per variable, EPV) [17]. Некоторые авторы описывают ситуации, когда требуется менее 10 EPP, другие — увеличивают минимальный порог до 15 или 50 EPP [18, 19]. Проблема заключается в том, что любое эмпирическое правило является слишком упрощенным решением, а требуемое количество пациентов в исследовании будет зависеть от многих сложных аспектов, в частности частоты исхода, распределения и значимости предикторов, количества событий для каждой категории переменных. В рамках эмпирического подхода можно использовать формулы: $N = 10 \times k / P$ и $N = 100 + 50 \times k$, где N — размер выборки, k — количество независимых переменных в финальной модели, P — частота изучаемого события [20, 21].

Не существует единого правила, основанного на EPP, которое гарантировало бы точную оценку параметров логистической регрессии. При использовании методов МО для каждого предиктора может потребоваться намного больше событий по сравнению с классическими методами моделирования, такими как логистическая регрессия, и даже при EPP = 200 модель может демонстрировать нестабильность и переобучение. Основная причина этого заключается в том, что количество параметров, учитываемых при МО, обычно намного превышает число исходных параметров из-за проверки множества условий их взаимодействий. Следовательно, методы МО не застрахованы от требований к размеру выборки и на самом деле нуждаются в действительно «больших данных», чтобы гарантировать корректный результат [17].

Кодирование предикторов и трансформация данных

Важной задачей при оценке данных перед созданием модели является их проверка на сбалансированность — распределение частот переменных. В условиях дисбаланса данных большинство моделей обычно чрезмерно предсказывает более крупные классы из-за их повышенной априорной вероятности, игнорируя меньшие с низкой частотой встречаемости. Часто редкий класс, например болезнь, является наиболее важным для точности прогноза. В этом случае помогают статистические методы устранения дисбаланса данных до этапа моделирования [22]. Несбалансированность данных по частоте встречаемости свойственна не только исходу, но и предикторам. Категории с низкой частотой встречаемости (предикторы с «почти нулевой дисперсией») объединяются с похожими другими с формированием новой переменной либо исключаются из анализа.

Числовые предикторы не должны подвергаться дихотомии (категоризации ниже и выше определенного предела) на этапе моделирования, поскольку происходит потеря ценной информации [12]. Конечный результат допустимо представлять в виде 2–4 категорий (групп риска), если потеря информации не станет при этом значительной [23].

Категориальные переменные с более двух значений рекомендуется трансформировать в фиктивные переменные, которые записываются в бинарном виде (1/0), а их общее число равно n-1, где n — количество исходных категорий. Например, переменная «стадия лимфомы по классификации Ann Arbor», имеющая 4 значения (I, II, III и IV), должна быть трансформирована в 3 фиктивные переменные: «стадия IV», «стадия III», «стадия II» со значениями 1 и 0, где 1 — пациенты с соответствующей стадией, 0 — все остальные. При этом «стадия I» не требует создания отдельной переменной, т. к. информация о ней закодирована в остальных предикторах.

Строго рекомендуемым допущением для применения некоторых методов моделирования, в частности МО, является математическая предобработка данных для приведения к определенному формату и представлению, что обеспечивает их корректное применение в многомерном анализе, который включает в себя удаление выбросов, логарифмирование, стандартизацию, степенное преобразование Бокса—Кокса и др. [24]. Часто используются методы уменьшения размерности исходных данных при большом числе предикторов, наиболее известный из них — метод главных компонент.

Выбор предикторов для модели

Изначально большой набор признаков создает избыточность информации («шум»), что нивелирует значимость важных факторов, особенно в условиях малого размера выборки. В связи с этим необходима селекция предикторов — сокращение исходно избыточного числа входных переменных без потери качества модели. Для решения этой задачи используются различные стратегии. Обычно при выборе предикторов в первую очередь следует полагаться на собственные знания об изучаемой проблеме и результаты предыдущих исследований в данной области, а затем — на методы статистического отбора. Метод пошагового регрессионного анализа имеет несколько недостатков. В частности, когда количество событий невелико и выборка недостаточно репрезентативна, расчетные коэффициенты регрессии нестабильны, а эффективность выбранной модели завышена [25]. Более сложные методы селекции предикторов, такие как рекурсивное исключение признаков, регрессия LASSO и другие методы МО, помогают создать модель, которая максимально соответствует исходным данным [26]. При выборе оптимальной модели из нескольких, различающихся между собой по числу переменных, предпочтение следует отдавать более простой, несмотря на незначительную потерю ее эффективности.

Моделирование и оценка качества модели

Выбор метода моделирования зависит от задачи, поставленной исследователем, типа данных и соблюдения допущений (условий) к его применению. Например, главным условием для применения широко распространенного регрессионного анализа Кокса является пропорциональность рисков: риски, связанные с переменной, не меняются с течением времени. Это условие требует обязательной проверки статистическими или графическими методами, а его отсутствие исключает правильную интерпретацию показателя отношения рисков [11].

Качество моделирования оценивается на исходных данных (тренировочный и валидационный датасеты). В зависимости от метода показателями эффективности могут быть такие, как чувствительность, специфичность, точность (accuracy/precision), F-мера, AUROC (с-индекс), критерий Акаике (AIC), среднеквадратическая ошибка (RMSE — стандартное отклонение остатков) [27]. Эти характеристики позволяют выбрать наилучшую модель в ходе экспериментов. Некоторые статистические методы позволяют интерпретировать полученные модели через коэффициенты регрессии и ОШ (логистическая регрессия, анализ Кокса) или графический анализ (дерево классификаций). Большинство МПМ, полученных с помощью методов МО, не поддается интерпретации (проблема «черного ящика»), и результат их работы воспринимается только через оценку эффективности.

Оценка производительности модели

Производительность модели — эффективность, проверяемая на внешних данных (валидационный и тестовый датасеты), не участвовавших в ее создании (обучении). Методы оценки производительности — дискриминативность и калибровка. Критерием дискриминативности модели является с-индекс, который визуализируется в виде ROC-кривой и оценивается количественно путем расчета площади под кривой (AUROC). Он представляет собой степень или вероятность, с которой модель может различать пациентов с исходом и без него. ROC-кривая строится в зависимости от показателей чувствительности (вертикальная ось) и специфичности (горизонтальная ось). AUROC не равняется точности модели, а представляет собой лишь вероятность, с которой она может работать (достоверность модели). Чем выше величина AUROC, тем выше ее гипотетическая точность; при значении 0,5 эффективность модели не отличается от случайного угадывания [28].

Истинная точность модели проверяется на тестовых и валидационных данных путем построения «матрицы путаницы» (confusion matrix) — таблицы производительности модели классификации на наборе тестовых данных, для которых известны истинные значения [29]. Матрица позволяет вычислить частоты истинно положительных (TP), истинно отрицательных (TN), ложноположительных (FP, ошибка первого рода) и ложноотрицательных (FN, ошибка второго рода) результатов. Исследователь также может рассчитать ряд дополнительных параметров.

Recall (чувствительность) — частота верно предсказанных TP-результатов из их общего числа.

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Precision (точность) — частота верно предсказанных TP-результатов из общего числа истинно и ложно предсказанных положительных результатов.

$$\text{Precision} = \frac{TP}{TP + FP}.$$

Accuracy (точность) — показатель, который описывает общую точность предсказания модели по всем классам. Это особенно полезно, когда каждый класс одинаково важен. Он рассчитывается как отношение количества правильных прогнозов к их общему количеству.

$$\text{Accuracy} = \frac{TP + FN}{TP + TN + FP + FN}.$$

F-score (F-мера) — показатель, который помогает комплексно оценивать Recall и Precision, используя среднее гармоническое вместо среднего арифметического.

$$\text{F-score} = \frac{2 \times Recall \times Precision}{Recall + Precision}.$$

Калибровка модели — процесс постобработки обученной модели, повышающий точность вычисляемой ею вероятности. Формально модель считается идеально откалиброванной, если для любого значения ее вероятности (P) достоверное предсказание класса верно в P × 100 % случаев. Прогнозируемые вероятности, которые соответствуют ожидаемому распределению вероятностей для каждого класса, называются калиброванными [30]. Проблема заключается в том, что не все модели МО способны предсказывать калиброванные вероятности. Некоторые алгоритмы, которые вычисляют истинное значение вероятности, часто не нуждаются в калибровке, например логистическая регрессия. Другие популярные модели, такие как деревья решений, метод опорных векторов, требуют калибровки своих аппроксимированных вероятностей.

Когда прогностические модели строятся на основе популяции, отличной от той, в которой они будут использоваться, их применение приводит к большим «остаткам» (ошибкам прогнозирования) из-за факторов, которые трудно учесть. Это становится причиной ошибочных решений в отношении конкретного пациента, даже если средний остаточный показатель для всей популяции очень низкий. Качественно откалиброванная модель минимизирует ошибки прогнозирования. Высокая точность модели на обучающем наборе данных необязательно означает ее эффективность из-за возможного наличия проблемы «переобучения». Не существует универсального и наилучшего метода калибровки; его выбор зависит от дизайна научного исследования и исходных данных [31].

Валидация модели

Существует внутренняя и внешняя валидация МПМ. Внутренняя валидация характеризует стабильность выбранных предикторов и качество прогнозов. Использование случайной тренировочной выборки для разработки модели и оставшихся пациентов для ее проверки является распространенным, но неоптимальным вариантом внутренней валидации. Лучшими методами считаются перекрестная проверка (кросс-валидация) и бутстрэппинг [2]. В основе метода перекрестной проверки лежит разделение исходного множества данных на k примерно равных блоков, например k = 10. Затем на 9 (k-1) блоках про-

водится обучение модели, а 10-й блок используется для тестирования. Процедура повторяется k раз, при этом на каждом проходе для проверки выбирается новый блок, а обучение проводится на оставшихся. Перекрестная проверка имеет важное преимущество перед применением одного множества для обучения и одного — для тестирования модели. Если при каждом проходе можно оценить выходную ошибку модели и усреднить ее по всем итерациям, то полученная оценка будет более достоверной. Бутстрэп-процедура состоит в том, чтобы случайным образом многократно извлекать повторные выборки из эмпирического распределения. Таким способом можно сформировать любое, сколь угодно большое, число бутстрэп-выборок (обычно 500–1000), каждая из которых содержит около $^{2}/_{3}$ уникальных значений начальной совокупности данных. В результате модификации частотного распределения исходных данных ожидается, что каждая следующая генерируемая псевдовыборка будет обладать значением оцениваемого параметра, немного отличающимся от первоначального. На основе разброса значений анализируемого показателя, полученного в ходе такой имитации, можно построить, например, его доверительные интервалы. Внутреннюю валидацию следует всегда проводить при разработке модели прогнозирования.

Внешняя валидация считается более строгим тестом для моделей прогнозирования, чем внутренняя, поскольку она касается применения модели на внешних данных, а не ее воспроизводимости. Внешняя валидация проводится путем изучения пациентов из других больниц (географическая валидация) или больных, которые недавно наблюдались либо будут наблюдаться (валидация во времени) [32]. Первоначальная точность модели сравнивается с таковой на всех этапах внутренней и внешней валидации.

Для исходов, возникающих во времени, таких как рецидив или смерть от лейкоза, смерть по другим причинам является конкурирующим риском. Показатели эффективности модели должны учитывать такие конкурирующие события. Отсутствие учета конкурирующих событий при разработке модели приводит к переоценке абсолютного риска, а их учета во время проверки — к искаженному представлению о производительности модели. При наличии конкурирующих рисков их необходимо учитывать при валидации модели [33, 34].

Презентация модели

На последнем этапе происходит представление прогностической модели в том виде, в котором она наилучшим образом соответствует потребностям реальной клинической практики [12]. Например, модель может быть презентована в упрощенном дизайне балльной шкалы для ручного подсчета и определения группы риска. В качестве баллов могут использоваться округленные показатели ОШ значимых предикторов. Графическим представлением регрессионных моделей являются номограммы. В последнее время модели прогнозирования часто публикуются в виде онлайн-калькуляторов или приложений для мобильных телефонов и планшетов, однако при этом сами расчеты риска остаются скрытыми. Модели на основе МО, как правило, представляют собой «черный ящик». МПМ вполне могут быть встроены в медицинские информационные системы врачей для поддержки принятия клинических решений.

ОСНОВНЫЕ ОШИБКИ ПРИ СОЗДАНИИ МПМ

Неопределенность цели создания МПМ

Иногда исследователь неправильно формулирует конечную цель научной работы — прогноз, объяснение причинно-следственной связи или описание данных. Путаница в этих понятиях приводит к ошибкам применения методов статистического анализа и их интерпретации. Залог успешного создания МПМ заключается не только и не столько в хороших данных и алгоритмах, сколько в знаниях в области медицины и статистики, а также умении правильно интерпретировать полученные результаты. Статистика не отвечает на вопрос «почему?» и не объясняет причинно-следственные связи, так же как методы описательной статистики и сравнения данных не предназначены для их прогнозирования.

Неэквивалентность статистической и клинической значимости

Интерпретация отсутствия статистической значимости ($p > 0,05$) как отсутствие практического эффекта является широко распространенной ошибкой. Значение $p > 5$ % ($p > 0,05$) считается «статистически незначимым». Под этим термином ошибочно подразумевается, что исследование показало отсутствие различий, тогда как на самом деле это отсутствие доказательств различия. Это совсем разные утверждения. Причиной получения статистически незначимых результатов часто бывает недостаточный размер выборки, что снижает мощность статистического теста для выявления реальных и клинически значимых различий [35]. При построении многопараметрических моделей методом логистической регрессии переменные, не имеющие статистической значимости, часто ошибочно исключаются из модели, снижая ее общую точность.

Интерпретация причинно-следственных связей

В медицинской науке существует множество примеров публикации противоречивых результатов о роли одних и тех же факторов в риске развития заболеваний, особенно в области эпидемиологии, а также между рандомизированными и обсервационными исследованиями [36]. Причиной этого являются систематические ошибки (от англ. *bias* — смещение) — неслучайное отклонение результатов и выводов от истины вследствие ошибок в дизайне и проведении исследования.

Интерпретация причинно-следственной взаимосвязи между данными только на основании $p < 0,05$ или тестах корреляции — частая ошибка. Всегда существует риск ложноположительного результата, снизить который можно путем применения более строгих уровней значимости для предикторов, используя в качестве критерия значимости $p \leq 0,001$, а не $p < 0,05$ [37, 38]. Кроме того, необходимо учитывать

риск предвзятости при выборе данных для анализа (например, если исследователь формирует группы сравнения, пренебрегая генератором случайных чисел). Получаемые в ходе статистического анализа причинно-следственные связи должны логически объясняться. Существуют ложные причинно-следственные взаимосвязи между данными. Например, уровни лейкоцитов и тромбоцитов при хроническом миелолейкозе могут сильно коррелировать между собой по причине прогрессирования заболевания.

Термин «фактор риска» означает причину изучаемого явления, предиктор (прогностический фактор, не являющийся причиной) или ковариату (переменную, статистически значимо связанную с исходом). Интерпретация полученной модели с объяснением входящих в нее переменных, когда это возможно, должна присутствовать в результатах научного исследования во избежание ложного заключения о причинно-следственных связях [32].

Пренебрежение ошибками в исходных данных

Качественные данные — основа эффективной МПМ. Часто наборы данных содержат ошибки, например, из-за неверного ввода, неточных измерений или опечаток в записях. Все это создает «информационный шум». Мнение о том, что только сильные и значимые предикторы обнаруживаются во время анализа, а остальные нивелируются, является заблуждением [39]. «Шумные данные» приводят к ошибкам измерений. Большой размер выборки улучшает, но не всегда исправляет ситуацию. Результаты измерений, выделяющиеся из общей выборки, в статистике называются «выбросами». Как правило, от «выбросов» избавляются, т. к. многие методы статистического анализа, в т. ч. МО, чувствительны к ним. В связи с этим крайне важно следить за качеством данных, проводить их описательно-статистический и графический анализ до этапа создания МПМ, а при необходимости выполнять их коррекцию или трансформацию.

Дихотомия

В медицинских исследованиях количественные переменные часто преобразуются в категориальные путем группировки значений по двум или более категориям. Данная процедура называется дихотомией. Дихотомия приводит к нескольким проблемам. Во-первых, большая часть информации теряется, поэтому статистическая мощность для обнаружения связи между переменной и исходом пациента снижается. Например, дихотомия переменной по медиане снижает мощность статистического теста на ту же величину, что и отбрасывание $^{1}/_{3}$ данных [40]. Во-вторых, можно серьезно недооценить степень различий в результатах между группами. Пациенты, близкие к точке отсечения, но находящиеся по разные стороны от нее, характеризуются как очень разные, являясь на самом деле очень похожими друг на друга. В-третьих, категоризация скрывает любую нелинейность в отношении между переменной и результатом. Кроме того, использование оптимальной точки отсечения, основанной на данных, ведет к серьезной систематической ошибке [41]. Использование оптимальной точки отсечения (например в ROC-анализе) приводит к искусственному полному или квазиполному разделению данных в выборке и увеличивает риск ложноположительного результата (ошибка первого рода) [42]. Дихотомия переменных негативно сказывается на результатах многофакторного анализа. Увеличивается вероятность ошибки второго рода и отказа от предиктора, который в реальности вносит большой вклад в объяснение зависимой переменной [43].

Недооценка влияния случайных эффектов

Широкое внедрение МПМ невозможно без внешней проверки на пациентах из других медицинских центров. Большие наборы новых данных помогают выяснить работоспособность модели, требуется ли ее обновление или адаптация, например повторная калибровка, или следует отказаться от ее использования. Иногда МПМ создаются в ходе многоцентровых исследований, тогда информация для их обучения берется из разных источников. С одной стороны, такой подход помогает экстраполировать результаты на большую часть изучаемой популяции, с другой — создать дополнительные методологические проблемы и трудности моделирования [44]. Примером является метаанализ, качество которого зависит от гетерогенности данных при его выполнении [45]. Гетерогенность, или неоднородность, данных, связанная с медицинским центром, влияет на качество прогноза модели, что называется случайным эффектом. При создании модели коэффициенты регрессии для фиксированных эффектов корректируются с учетом случайных эффектов (метод смешанной регрессии). Факторы, вызывающие случайный эффект, могут использоваться в качестве дополнительных независимых переменных при моделировании.

Проблема размера выборки

Большие размеры выборки позволяют разрабатывать более надежные модели. Данные должны быть определенного качества и репрезентативными для поставленных задач и условий применения. Существуют как простые эмпирические, так и более сложные современные методы оценки размера выборки для разных типов исходов, направленные на достижение баланса между переобучением модели и минимизацией ложноотрицательных результатов, описанные R.D. Riley и соавт. [46–49].

Игнорирование отсутствующих данных

Пропуски в данных неизбежны в эпидемиологических и клинических исследованиях, но их способность снизить достоверность результатов часто упускается из виду. Устранить переменные с пропущенными значениями можно, включая в анализ только тех пациентов, у которых нет недостающей информации (полный анализ случаев). Однако при этом результаты анализа будут необъективными. Кумулятивный эффект отсутствующих данных по нескольким переменным часто приводит к исключению значительной части исходной выборки, что, в свою очередь, вызывает существенную потерю точности и мощности статистического теста [50]. Перспективным решением этой проблемы является множественное вменение пустых значений (замена пустых значений правдо-

подобными). Самый простой, но самый неточный метод — наивный подход, который предполагает замену пустых значений средними показателями (арифметическим средним, медианой). Более эффективными решениями считаются методы вменения на базе МО (регрессия, случайный лес, метод ближайшего соседа и др.) [51]. Вменение пропущенных значений предпочтительнее полного анализа случаев при условии правильно выбранного метода. Характер и частота пропущенных значений, причины пропуска и метод вменения должны быть отражены исследователем при описании процесса разработки МПМ в научной публикации или технической документации.

Выбор предикторов только на основе анализа данных

Проблема выбора предикторов в многопараметрических регрессионных моделях возникает в процессе поиска факторов, влияющих на целевой результат. Существует большое количество решений, но ни одно из них не является идеальным. Прежде чем использовать статистические методы выбора переменных, следует решить, нужны ли такие методы в конкретном исследовании. При небольшом числе переменных можно использовать эмпирический выбор на основе предыдущих исследований в этой области. Если же автоматическая селекция предикторов необходима, то следует выбирать оптимальный из существующих доступных методов. Например, популярная рекомендация по выполнению однофакторного анализа как предварительного этапа перед многофакторным является не самым эффективным вариантом [52]. Альтернативой может быть обратная пошаговая регрессия с использованием в качестве критерия селекции модели $p = 0{,}157$ или AIC, особенно для малых по размеру выборок с EPP < 100 [53]. К современным подходам к поиску стабильных предикторов относятся регрессия LASSO и методы МО [54].

Игнорирование тестовой выборки

Часто исследователи, построив прогностическую модель, делают выводы о ее высокой эффективности на тех же данных, не применяя валидацию на тестовой выборке [55]. Использование тестового набора данных, который не применяли в обучении модели, является обязательным условием проверки ее качества. Результаты работы модели при этом могут стать хуже, а при ее переобучении и вовсе не отличаться от случайного предсказания. Тем не менее это будет более приближенный к реальности результат. Сопоставимая точность модели на тренировочном и тестовом наборах данных означает хорошее качество ее обучения.

Плохая отчетность

Качество описания всех этапов создания прогностической модели в публикациях остается низким. Только при наличии полной информации о всех аспектах моделирования можно адекватно оценить риск ошибок и потенциальную полезность МПМ. Для внутреннего и внешнего контроля за процессом разработки МПМ рекомендуется пользоваться руководствами по составлению прозрачной отчетности TRIPOD, оценке риска систематической ошибки и применимости моделей прогноза PROBAST, а также библиотекой отчетов EQUATOR, доступной на сайте: equator-network.org [14, 56]. В этих руководствах и отчетах содержится список из десятков пунктов и подробных комментариев к ним, целью которых является улучшение отчетности и повышение доверия к исследованиям, связанным с разработкой ДММ и ПММ.

ЗАКЛЮЧЕНИЕ

В статье описаны основные обязательные шаги, необходимые при разработке МПМ. Существует большое число деталей, которые необходимо учитывать для каждого этапа разработки и проверки модели, описанных в соответствующей методологической литературе. Для создания качественной МПМ обычно требуется привлечение экспертов по статистике. Частыми ошибками при разработке МПМ бывают следующие: несоблюдение условий, необходимых для применения статистического метода моделирования, отсутствие проверки эффективности модели на тестовой выборке, неполная отчетность, использование нерепрезентативных данных, моделирование на выборке малого размера, дихотомия переменных, отсутствие калибровки, внутренней или внешней валидации.

Внедрение статистической МПМ в качестве основы для системы поддержки принятия врачебных решений требует ряда дополнительных шагов. Модель должна быть тщательно проверена на внешних репрезентативных данных (внешняя валидация) в рамках клинических исследований. Разработчикам настоятельно рекомендуется четко описывать любые различия между данными и результатами, используемыми и полученными при создании модели и при ее проверке, там где это возможно. Если МПМ создается с использованием технологий МО, необходимо следовать рекомендациям, изложенным в руководстве по составлению отчетов для ранней клинической оценки систем поддержки принятия решений, управляемых искусственным интеллектом (DECIDE-AI) [57]. В России разрабатываются отечественные ГОСТы в области искусственного интеллекта.

Конечным этапом внедрения МПМ в клиническую практику является разработка программного обеспечения для взаимодействия пользователей через существующую цифровую инфраструктуру (медицинские и лабораторные информационные системы, медицинские калькуляторы, веб- и мобильные приложения). Особое внимание следует уделить дизайну продукта, удобству его использования в повседневной работе и безопасности обмена данными.

Внедрение МПМ в практическое здравоохранение вне клинических исследований включает в себя обучение пользователей взаимодействию с продуктом и интерпретации результатов, обслуживание и обновление, мониторинг работы системы в реальной практике и аудит. Необходимо руководство по долгосрочным аспектам использования модели, повторной калибровке (обновлению) и периодическому пере-

обучению модели при необходимости с указанием календарных сроков.

Изложенные методологические подходы целесообразно применять при разработке и проверке моделей прогнозирования. На этой основе следует проводить систематический анализ и критическую оценку публикаций, посвященных МПМ. Следование данным рекомендациям повысит эффективность создаваемых моделей прогнозирования в медицине.

КОНФЛИКТЫ ИНТЕРЕСОВ

Автор заявляет об отсутствии конфликтов интересов.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Исследование не имело спонсорской поддержки.

ЛИТЕРАТУРА/REFERENCES

1. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. Br Med J. 2020;369:m1328. doi: 10.1136/bmj.m1328.

2. Van Smeden M, Reitsma JB, Riley RD, et al. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. 2021;132:142–5. doi: 10.1016/j.jclinepi.2021.01.009.

3. Schalling M, Gleiss A, Gisslinger B, et al. Essential thrombocythemia vs. pre-fibrotic/early primary myelofibrosis: discrimination by laboratory and clinical data. Blood Cancer J. 2017;7(12):643. doi: 10.1038/s41408-017-0006-y.

4. Guncar G, Kukar M, Notar M, et al. An application of machine learning to haematological diagnosis. Sci Rep. 2018;8(1):411. doi: 10.1038/s41598-017-18564-8.

5. Sehn LH, Berry B, Chhanabhai M, et al. The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with diffuse large B-cell lymphoma treated with R-CHOP. Blood. 2007;109(5):1857–61. doi: 10.1182/blood-2006-08-038257.

6. Van de Schans SAM, Steyerberg EW, Nijziel MR, et al. Validation, revision and extension of the Follicular Lymphoma International Prognostic Index (FLIPI) in a population-based setting. Ann Oncol. 2009;20(10):1697–702. doi: 10.1093/annonc/mdp053.

7. Palumbo A, Avet-Loiseau H, Oliva S, et al. Revised International Staging System for Multiple Myeloma: A Report From International Myeloma Working Group. J Clin Oncol. 2015;33(26):2863–9. doi: 10.1200/JCO.2015.61.2267.

8. Лучинин А.С. Искусственный интеллект в гематологии. Клиническая онкогематология. 2022;15(1):16–27. doi: 10.21320/2500-2139-2022-15-1-16-27.

[Luchinin AS. Artificial Intelligence in Hematology. Clinical oncohematology. 2022;15(1):16–27. doi: 10.21320/2500-2139-2022-15-1-16-27. (In Russ)]

9. Zhou L, Meng X, Huang Y, et al. An interpretable deep learning workflow for discovering subvisual abnormalities in CT scans of COVID-19 inpatients and survivors. Nat Mach Intell. 2022;4(5):494–503. doi: 10.1038/s42256-022-00483-7.

10. Szumilas M. Explaining Odds Ratios. J Can Acad Child Adolesc Psychiatry. 2010;19(3):227–29.

11. Barraclough H, Simms L, Govindan R. Biostatistics Primer: What a Clinician Ought to Know: Hazard Ratios. J Thorac Oncol. 2011;6(6):978–82. doi: 10.1097/JTO.0b013e31821b10ab.

12. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925–31. doi: 10.1093/eurheartj/ehu207.

13. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi: 10.1186/s12916-019-1466-7.

14. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170(1):51–8. doi: 10.7326/M18-1376.

15. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. Br Med J. 2009;338:b606. doi: 10.1136/bmj.b606.

16. Altman DG, Bland JM. Missing data. Br Med J. 2007;334(7590):424. doi: 10.1136/bmj.38977.682025.2C.

17. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. Br Med J. 2020;368:m441. doi: 10.1136/bmj.m441.

18. Jenkins DG, Quintana-Ascencio PF. A solution to minimum sample size for regressions. PloS One. 2020;15(2):e0229345. doi: 10.1371/journal.pone.0229345.

19. Van Voorhis WCR, Morgan BL. Understanding Power and Rules of Thumb for Determining Sample Sizes. Tutor Quant Meth Psychol. 2007;3(2):43–50. doi: 10.20982/tqmp.03.2.p043.

20. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9. doi: 10.1016/s0895-4356(96)00236-3.

21. Bujang MA, Sa'at N, Sidik TMITAB, Joo LC. Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. Malays J Med Sci. 2018;25(4):122–30. doi: 10.21315/mjms2018.25.4.12.

22. Zhou P-Y, Wong AKC. Explanation and prediction of clinical data with imbalanced class distribution based on pattern discovery and disentanglement. BMC Med Inform Decis Mak. 2021;21(1):16. doi: 10.1186/s12911-020-01356-y.

23. Pauker SG, Kassirer JP. The Threshold Approach to Clinical Decision Making. N Engl J Med. 1980;302(20):1109–17. doi: 10.1056/NEJM198005153022003.

24. Lee DK. Data transformation: a focus on the interpretation. Korean J Anesthesiol. 2020;73(6):503–8. doi: 10.4097/kja.20137.

25. Zhang Z. Variable selection with stepwise and best subset approaches. Ann Transl Med. 2016;4(7):136. doi: 10.21037/atm.2016.03.35.

26. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.

27. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med. 2022;5(1):1–13. doi: 10.1038/s41746-021-00549-7.

28. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. 2013;4(2):627–35.

29. Agarwal A, Sharma P, Alshehri M, et al. Classification model for accuracy and intrusion detection using machine learning approach. PeerJ Comput Sci. 2021;7:e437. doi: 10.7717/peerj-cs.437.

30. Hendriksen JMT, Geersing GJ, Moons KGM, de Groot JAH. Diagnostic and prognostic prediction models. J Thromb Haemost. 2013;11(Suppl 1):129–41. doi: 10.1111/jth.12262.

31. Huang Y, Li W, Macheret F, et al. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc. 2020;27(4):621–33. doi: 10.1093/jamia/ocz228.

32. Snell KIE, Archer L, Ensor J, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. J Clin Epidemiol. 2021;135:79–89. doi: 10.1016/j.jclinepi.2021.02.011.

33. Ramspek CL, Teece L, Snell KIE, et al. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. Int J Epidemiol. 2022;51(2):615–25. doi: 10.1093/ije/dyab256.

34. Van Geloven N, Giardiello D, Bonneville EF, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. Br Med J. 2022;377:e069249. doi: 10.1136/bmj-2021-069249.

35. Altman DG, Bland JM. Absence of evidence is not evidence of absence. Br Med J. 1995;311(7003):485. doi: 10.1136/bmj.311.7003.485.

36. Smith GD, Ebrahim S. Data dredging, bias, or confounding. Br Med J. 2002;325(7378):1437–8. doi: 10.1136/bmj.325.7378.1437.

37. Lakens D, Adolfi FG, Albers CJ, et al. Justify your alpha. Nat Hum Behav. 2018;2(3):168–71. doi: 10.1038/s41562-018-0311-x.

38. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. Nat Hum Behav. 2018;2(1):6–10. doi: 10.1038/s41562-017-0189-z.

39. Van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. Int J Epidemiol. 2020;49(1):338–47. doi: 10.1093/ije/dyz251.

40. Altman DG, Royston P. The cost of dichotomising continuous variables. Br Med J. 2006;332(7549):1080. doi: 10.1136/bmj.332.7549.1080.

41. Wynants L, van Smeden M, McLernon DJ, et al. Three myths about risk thresholds for prediction models. BMC Med. 2019;17(1):192. doi: 10.1186/s12916-019-1425-3.

42. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127–41. doi: 10.1002/sim.2331.

43. Vargha A, Rudas T, Delaney HD, Maxwell SE. Dichotomization, Partial Correlation, and Conditional Independence. J Educ Behav Stat. 1996;21(3):264–82. doi: 10.3102/10769986021003264.

44. Basagana X, Pedersen M, Barrera-Gomez J, et al. Analysis of multicentre epidemiological studies: contrasting fixed or random effects modelling and meta-analysis. Int J Epidemiol. 2018;47(4):1343–54. doi: 10.1093/ije/dyy117.

45. Лучинин А.С. Лечение пациентов с впервые диагностированной диффузной В-крупноклеточной лимфомой: обзор литературы и метаанализ. Клиническая онкогематология. 2022;15(2):130–9. doi: 10.21320/2500-2139-2022-15-2-130-139.

[Luchinin AS. Treatment of Patients with Newly Diagnosed Diffuse Large B-Cell Lymphoma: A Literature Review and Meta-Analysis. Clinical oncohematology. 2022;15(2):130–9. doi: 10.21320/2500-2139-2022-15-2-130-139. (In Russ)]

46. Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. Stat Med. 2022;41(7):1280–95. doi: 10.1002/sim.9275.

47. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part I – continuous outcomes. Stat Med. 2019;38(7):1262–75. doi: 10.1002/sim.7993.

48. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: Part II – binary and time-to-event outcomes. Stat Med. 2019;38(7):1276–96. doi: 10.1002/sim.7992.

49. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med. 2021;40(19):4230–51. doi: 10.1002/sim.9025.

50. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Br Med J. 2009;338:b2393. doi: 10.1136/bmj.b2393.

51. Petrazzini BO, Naya H, Lopez-Bello F, et al. Evaluation of different approaches for missing data imputation on features associated to genomic data. BioData Min. 2021;14(1):44. doi: 10.1186/s13040-021-00274-7.

52. 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(8):907–16. doi: 10.1016/0895-4356(96)00025-x.

53. Heinze G, Dunkler D. Five myths about variable selection. Transpl Int. 2017;30(1):6–10. doi: 10.1111/tri.12895.

54. Chen R-C, Dewi C, Huang S-W, Caraka RE. Selecting critical features for data classification based on machine learning methods. J Big Data. 2020;7(1):52. doi: 10.1186/s40537-020-00327-4.

55. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691–8. doi: 10.1136/heartjnl-2011-301247.

56. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73. doi: 10.7326/M14-0698.

57. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924–33. doi: 10.1038/s41591-022-01772-9.