DOI: 10.28996/2618-9801-2022-1-99-113

«Подводные камни» статистического анализа и клинической интерпретации полученных оценок на примере пациентов с заболеваниями почек

Часть IV: ROC-анализ и специальные показатели информативности биомаркеров

А.Б. Зулькарнаев[1]*, Е.В. Паршина[2], В.А. Федулкина[1]

[1] Хирургическое отделение трансплантации почки, ГБУЗ МО «Московский областной научно-исследовательский клинический институт им. М.Ф. Владимирского», 129110, Москва, ул. Щепкина, 61/2, корп. 6, Российская Федерация

[2] Отделение амбулаторного диализа, Санкт-Петербургский государственный университет, Клиника высоких медицинских технологий им. Н.И. Пирогова, 198103, Санкт-Петербург, набережная реки Фонтанки, 154, Российская Федерация

Для цитирования: Зулькарнаев А.Б., Паршина Е.В., Федулкина В.А. «Подводные камни» статистического анализа и клинической интерпретации полученных оценок на примере пациентов с заболеваниями почек. Часть IV: ROC-анализ и специальные показатели информативности биомаркеров. Нефрология и диализ. 2022; 24(1):99-113. doi: 10.28996/2618-9801-2022-1-99-113

Pitfalls of statistical analysis and clinical interpretation of the obtained estimates on the example of patients with kidney disease

Part IV: ROC analysis and special assessments of biomarker informativeness

A.B. Zulkarnaev[1]*, E.V. Parshina[2], V.A. Fedulkina[1]

[1] Surgical Department of Transplantology and dialysis, Moscow Regional Research and Clinical Institute, 61/2 Shchepkina str., Moscow, 129110, Russian Federation

[2] Department of outpatient dialysis, Saint Petersburg State University Hospital, 154 Fontanka emb., Saint-Petersburg, 198103, Russian Federation

For citation: Zulkarnaev A.B., Parshina E.V., Fedulkina V.A. Pitfalls of statistical analysis and clinical interpretation of the obtained estimates on the example of patients with kidney disease. Part IV: ROC analysis and special assessments of biomarker informativeness. Nephrology and Dialysis. 2022; 24(1):99-113. doi: 10.28996/2618-9801-2022-1-99-113

Ключевые слова: *ROC-анализ, коэффициент корреляции Метьюса, F-мера, чувствительность, специфичность, площадь под ROC-кривой, пороговое значение маркера, статистика*

Резюме

В настоящее время множество показателей претендуют на роль маркеров, позволяющих выявлять заболевания (скрининговые маркеры) или подтверждать заболевание (диагностические или прогностические маркеры). Одним из показателей, который часто применяется для оценки эффективности диагностического теста является Acc («Accuracy», «точность»), которая представляет собой долю верных классификаций.

Адрес для переписки: *Зулькарнаев Алексей Батыргараевич*
e-mail: *7059899@gmail.com*

Corresponding author: *Dr. Alexey B. Zulkarnaev*
e-mail: *7059899@gmail.com*

* *ORCID: 0000-0001-5405-7887*

Несмотря на то, что Acc часто приводится в публикациях как мера эффективности теста, она таковой не является. Более того, Acc может достигать больших значений даже при полном отсутствии реальной сопряженности маркера и исхода. Более стабильной оценкой является коэффициент корреляции Метьюса – MCC *(Matthews correlation coefficient).*

Другой интересной оценкой являются F-меры и, в частности, самая распространенная из них – F1. Показатель F1 представляет собой сбалансированную обобщенную оценку (гармоническое среднее) чувствительности (или «recall») и прогностической ценности положительного результата (или «precision»). Этот показатель позволяет более полно оценить способность теста распознавать пациентов с болезнью, но не отличать больных от здоровых, поскольку от не учитывает истинно отрицательные результаты.

В случае, когда маркер представляет собой не номинальный бинарный признак, а непрерывный количественный, бывает важно выявить порог, который позволит наиболее эффективно решать определенные задачи при помощи теста (относить субъектов к больным или здоровым на основании значения маркера). Традиционно для этой задачи используют ROC-анализ, выбирая оптимальное пороговое значение количественного признака на основании индекса Юдена (максимального расстояния от диагональной опорной линии на графике ROC-кривой) или *K*-индекса (минимального расстояния от ROC-кривой до левого верхнего угла графика). Такой утилитарный подход применим, когда пороговое значение обеспечивает большие значения чувствительности и специфичности (более 0,9). В большинстве случаев пороговое значение выбирается на основании максимизации (или достижения минимально приемлемого значения) определенных оценок: чувствительности, специфичности, положительной или отрицательной значимости, относительного риска или отношения шансов, отношения правдоподобия и др., что позволяет адаптировать маркер под определенные задачи.

Abstract

Currently, many classifiers claim to be markers that enable to detect (screening markers) or confirm a disease (diagnostic or prognostic markers). Accuracy (Acc) is a metric that is often used to evaluate the effectiveness of a diagnostic test, representing the proportion of correct classifications.

Although Acc is widely used in publications as a measure of test effectiveness, in fact, it isn`t so. Moreover, Acc can reach large values even if a marker and an outcome are completely not conjugated. A more balanced estimate is the Matthews correlation coefficient (MCC).

Another interesting evaluation metric is F-measure, in particular – the traditional F1-score. The F1-measure is a balanced average (harmonic mean) of sensitivity (or "recall") and positive predictive value (or "precision"). This metric allows us to more fully assess the ability of the test to recognize patients with the disease, but not to discriminate between sick and healthy subjects, since it does not consider true negative results.

In the case when the marker is not binary, but a continuous quantitative variable, it is important to identify a cut-off threshold that allows us to solve certain tasks in a more effective way using the test (to classify subjects as sick or healthy based on the marker value). Traditionally, ROC analysis is used for this purpose, choosing the optimal threshold value of a quantitative variable based on the Yuden index (the maximum distance from the diagonal reference line on the ROC curve graph) or the K-index (the minimum distance from the ROC curve to the upper left corner of the graph). Such a utilitarian approach is applicable when the threshold provides high values of both sensitivity and specificity (more than 0.9). In most cases, the threshold is chosen based on the maximization (or achievement of the minimum acceptable value) of certain estimates, such as sensitivity, specificity, positive or negative predictive value, relative risk or odds ratio, likelihood ratio, etc., which allows using the marker to carry out certain tasks.

Key words: *ROC analysis, Matthews correlation coefficient, F-measure, sensitivity, specificity, area under the ROC curve, classifier threshold value, statistics*

Введение

Практикующий доктор неизбежно вынужден принимать решения в условиях частичной неопределённости. В отличие от свойственной каждому специалисту интуиции, рациональный подход опирается на определенные оценки, которые позволяют измерить обоснованность принимаемых клинических решений, субъективное сделать объективным.

Напомним [1, 2], что в качестве основных мер сопряженности маркера и исхода используют отношение шансов – *odds ratio, OR* или отношение рисков – *risk ratio, RR*. Известны показатели, характеризующие исключительно скрининговую эффективность маркера: чувствительность (*sensitivity – Se*)

и специфичность (*specificity – Sp*), которые представляют собой долю носителей маркера среди больных и долю лиц, свободных от маркера, среди здоровых (иными словами – долю больных и здоровых, которые могут быть верно идентифицированы на основе значения маркера соответственно). Кроме этого известны показатели прогностической (диагностической) информативности маркера: оценки прогностической значимости положительного и отрицательного результата теста (*positive predictive value – PPV* и *negative predictive value – NPV*), которые представляют собой доли верно идентифицированных носителей маркера и лиц, свободных от маркера (или в более привычной вероятностной интерпретации – вероятность заболевания при наличии маркера и вероятность отсутствия заболевания при отсутствии маркера, соответственно). Также известны интегральные показатели информативности маркеров: площадь под (*ROC*) кривой (*area under the curve – AUC*) или скрининговая балансовая точность (*screening balance accuracy – SBA*) и прогностическая балансовая точность (*predictive balance accuracy – PBA*), которые представляют собой арифметическое среднее *Se* и *Sp*, *PPV* и *NPV*, соответственно. Альтернативными интегральными оценками скрининговой и прогностической ценности маркеров являются индекс Юдена (Youden's j statistic – $J_{Se\text{-}Sp}$) и похожий показатель – $J_{PPN\text{-}NPV}$ (этот показатель не имеет устойчивого названия), которые представляют собой разность между долей носителей маркера среди больных и лиц с маркером среди здоровых, разность между долей больных среди носителей маркера и больных среди лиц без маркера, соответственно. Все эти показатели рассчитываются на основании таблицы сопряженности два на два (в случае бинарного маркера и бинарного исхода – есть или нет заболевание). Напомним, что ячейки этой таблицы можно описать как:

- истинно положительные результаты: *true positive, TP:* маркер есть (*M*), заболевание есть (*D*);
- истинно отрицательные результаты – *true negative, TN:* маркера нет ($\overline{M}$), заболевания нет ($\overline{D}$);
- ложно положительные результаты – *false positive, FP:* маркер есть (M), заболевания нет ($\overline{D}$);
- ложно отрицательные результаты – *false negative, FN:* маркера нет ($\overline{M}$), заболевание есть (*D*).

Однако перечень оценок, которые можно получить из четырехпольной таблицы сопряженности, этим не ограничивается. В профильных публикациях нередко приводятся и иные показатели: точность (*accuracy – Acc*), коэффициент корреляции Метьюса, *F-меры*. Кроме этого, мы рассмотрим некоторые особенности ROC-анализа.

Мы остановимся на нескольких актуальных вопросах, ответ на которые не так очевиден, как принято считать.

1. Почему при оценке эффективности маркера нельзя просто определить долю верных классификаций: суммы частот больных субъектов с маркером (*TP*) и здоровых субъектов, свободных от маркера (*TN*), по отношению ко всей совокупности срабатываний теста (*TP+FP+FN+TN*)? Иными словами – вероятность верного срабатывания теста. Потому что этот показатель (*Acc*) может при определенных условиях (и не таких уж невероятных) принимать большие значения даже при полном отсутствии значимой сопряженности маркера и заболевания.
2. Как можно оценить эффективность маркера с учетом его и скрининговой (то есть способности эффективно выявлять больных при скрининге), и прогностической информативности (то есть с большой вероятностью свидетельствовать в пользу наличия заболевания при наличии маркера). Для этого существуют специальный показатель – *F-меры*.
3. Если маркер представляет собой не качественный, а количественный показатель, как выбрать оптимальное пороговое значение? Вопреки распространённому принципу выбора точки на *ROC*-кривой, имеющей минимальное расстояние «от левого верхнего угла» графика есть и иные принципы выбора оптимального порогового значения количественного признака.

1. Показатель «Accuracy» как альтернативная мера сопряженности маркера и заболевания. В дополнение к площади под ROC-кривой (*AUC*) и прогностической балансовой точности (*PBA*), интуитивно привлекательным является показатель, именуемый чаще всего «точность» (*accuracy – Acc*), который представляет собой долю верно идентифицированных лиц (*TP* и *TN*) в результате применения теста:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Считаем необходимым отметить, что в случае, когда статистические термины переводятся на русский язык, может возникнуть путаница. В частности, под термином «точность» нередко понимают не только *Acc*, но и показатель *«precision»* – «прецизионность» (он будет рассмотрен нами ниже), который обладает отличной от *Acc* интерпретацией. А именно: термином *«precision»* именуют *PPV* при вычислении *F-мер*. По нашему глубочайшему убеждению, когда это возможно, следует приводить название статистических показателей на английском языке. Наглядным примером является термин «отношение рисков», под которым может скрываться не только *«risk ratio»,* но и, например, *«hazard ratio»* или *«incidence rate ratio»,* а нередко авторы ошибочно переводят так термин *«odds ratio»*. Указание термина на английском языке вносит полную ясность и помогает читателю верно интерпретировать приводимые оценки.

Показатель *Acc* является легко интерпретируемым (доля верных срабатываний теста). Поскольку

Acc представляет собой вероятность того, что тест даст верный результат (но не более того), диапазон возможных значений *Acc* лежит в интервале от нуля до единицы. Значения $Acc \to 0$ (стремящиеся к нулю, малые значения) свидетельствует о малой частоте верных «срабатываний теста», $Acc \to 1$ – о большой частоте (при этом под верным «срабатыванием» теста нужно понимать отсутствие болезни при отсутствии маркера и наличие болезни при наличии маркера).

При всей интуитивной привлекательности этой оценки она таит в себе некоторые «подводные камни». А именно – как это ни парадоксально, *Acc* не является показателем эффективности теста и не отражает сопряженности маркера и заболевания.

Рассмотрим гипотетический, но наглядный *пример 1:*

	D	$\overline{D}$	*Σ*
M	1	5	6
$\overline{M}$	5	89	94
Σ	6	94	100

P_M=0,06 (частота встречаемости маркера); P_D=0,06 (распространенность заболевания); *RR*=3,13; *OR*=3,56; *Acc*=0,9; *Se*=0,167; *Sp*=0,947; *AUC*=0,557; *PPV*=0,167; *NPV*=0,947; *PBA*=0,557

	D	$\overline{D}$	*Σ*
M	19	36	55
$\overline{M}$	1	44	45
Σ	20	80	100

P_M=0,55; P_D=0,2; *RR*=15,6; *OR*=23,2; *Acc*=0,63; *Se*=0,95; *Sp*=0,55; *AUC*=0,75; *PPV*=0,346; *NPV*=0,978; *PBA*=0,662

Как видно из левой матрицы (где представлены уже знакомые вам [1, 2] показатели) маркер обладает низкой скрининговой и прогностической эффективностью, о чем свидетельствуют значения *AUC* и *PBA*, маркер слабо сопряжен с исходом, о чем свидетельствует значения *RR* и *OR*. Несмотря на то, что в этих работах мы старательно избегаем рассмотрения вопроса статистической значимости различий, отметим, что тут различия статистически не значимы: p=0,317 (точный тест Фишера). Тем не менее, значение *Acc*=0,9 (близко к максимальному). Контрпример дает нам матрицу справа: тест обладает неплохой скрининговой эффективностью (*AUC*=0,75), маркер сильно сопряжен с заболеванием (*RR*≈15,6; *OR*≈23,2). При этом *Acc*=0,63, а p<0,0001.

Дело в том, что, помимо силы сопряжённости маркера и заболевания, на *Acc* оказывают непосредственное влияние частота встречаемости маркера (P_M) и распространенность заболевания (P_D) (рисунок 1А) и в двух случаях это показатель может принимать большие значения даже в условиях отсутствия реальной сопряженности маркера и исхода. В первом, крайне маловероятном случае, когда и маркер, и заболевания встречаются очень часто ($P_M \to 1$ и $P_D \to 1$) $Acc \to 1$ даже при *RR*, $OR \to 1$ (*Acc* может достигать больших значений даже в условиях отсутствия сопряженности маркера и заболевания). Это вполне закономерно: почти все протестированные на самом деле будут больны и при этом иметь маркер, соответственно тест будет давать верное «срабатывание».

Во втором, противоположном и значительно более вероятном случае, *Acc* может достигать больших значений при очень редком заболевании и очень редко встречающемся маркере (даже при *RR*, $OR \to 1$). Читателю необходимо об этом помнить.

Закономерно, что при $P_M \to 0$ и $P_D \to 0$, $TP \to 0$, $FP \to 0$, $FN \to 0$, а $TN \to 1$. То есть больных носителей маркера будет очень мало, т.к. и заболевание, и маркер встречаются очень редко, при этом почти все субъекты будут здоровы и будут свободны от маркера и тест будет давать большую долю верных классификаций вне зависимости от того, сопряжен маркер с исходом или нет, применим маркер на практике или нет. В результате, как нетрудно заметить из формулы вычисления этого показателя, представленной выше, $Acc \to 1$ и слабо зависит от *OR*. Необходимо помнить, что показатель *Acc* может давать извра-

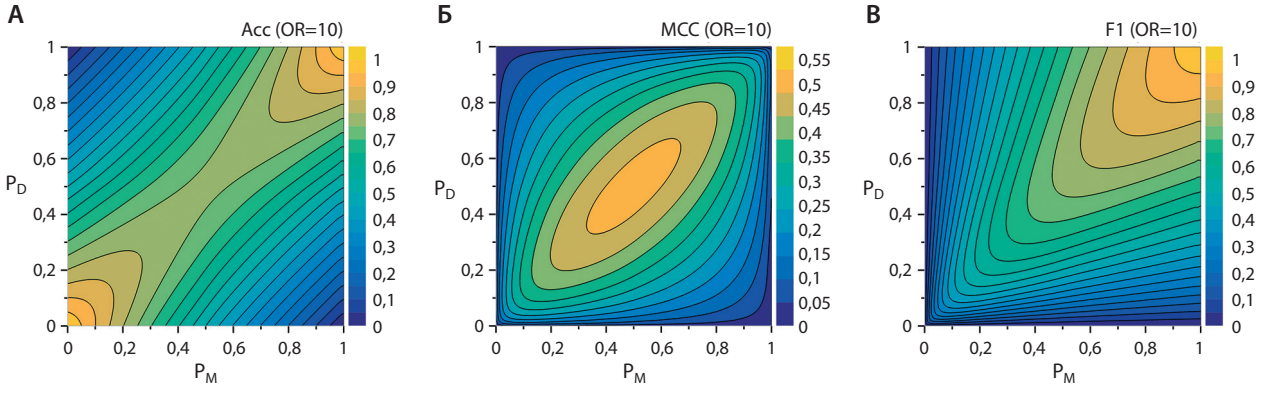

Рис. 1. Показатели информативности маркера при различных значениях распространенности заболевания (P_D) и частоты встречаемости маркера (P_M) при *OR*=10. Искусственный датасет.

Fig. 1. Markers performance estimates and outcome at different values of the disease prevalence (P_D) and the frequency of marker occurrence (P_M) when *OR*=10. Artificial dataset.

щенное представление о сопряженности маркера и исхода в условиях сильно несбалансированной матрицы: доля пациентов с заболеванием и маркером очень мала или очень велика.

Приведем еще два примера – опубликованные работы, где указано значение *Acc*.

Пример 2. Авторы оценили информативность нескольких показателей при диагностике диабетической нефропатии у пациентов с инсулиннезависимым сахарным диабетом [3]. Одним из наиболее информативных показателей была микроальбуминурия, которая обеспечивала *Acc*=0,86. Используя данные, приведенные в статье, восстановим исходную матрицу.

	D	$\overline{D}$	*Σ*
M	5	10	15
$\overline{M}$	9	114	123
Σ	14	124	138

P_M=0,11; P_D=0,11;
RR=4,56; *OR*=6,33; *Acc*=0,862;
Se=0,357; *Sp*=0,919; *AUC*=0,638;
PPV=0,333; *NPV*=0,927; *PBA*=0,63

Из примера видно, что несмотря на то, что развитие микроальбуминурии достаточно сильно сопряжено с риском развития нефропатии (*OR, RR,* χ^2 с поправкой Йейтса p=0,007, точный тест Фишера p=0,0086), маркер обладает весьма посредственными интегральными скрининговым и прогностическим показателями информативности (*AUC* и *PBA* соответственно) и в клинической практике будет практически бесполезен (разумеется, в тех условиях формирования выборки и оценки параметров, которые наблюдались в статье).

Более наглядным является *пример 3*, где авторы оценили информативность неселективной цифровой субтракционной ангиографии в идентификации раннего ветвления основной почечной артерии у прижизненных доноров почки [4]. Из данных, приведенных в статье, мы восстановили исходную матрицу:

	D	$\overline{D}$	*Σ*
M	2	2	4
$\overline{M}$	2	60	62
Σ	4	62	66

P_M=0,06; P_D=0,06;
RR=15,5; *OR*=30; *Acc*=0,939;
Se=0,5; *Sp*=0,968; *AUC*=0,734;
PPV=0,5; *NPV*=0,968; *PBA*=0,734

Видно, значение интегральных показателей и скрининговой (*AUC*), и прогностической (*PBA*) информативности маркера чуть больше минимально приемлемых для возможности практического использования (>0,7). В то время как показатель *Acc* свидетельствует о высокой информативности (0,939). Это наблюдается благодаря тому, что P_M и P_D имеют малые значения (0,06).

Мы не будем приводить исходную матрицу, но отметим только, что в другой статье, где оценивалась информативность различных вариантов перекрестной пробы у реципиентов почечного аллотрансплантата («классический» тест комплемент-зависимой цитотоксичности, проточная цитофлуориметрия, мультиплексный анализ на платформе «Luminex») [5], хорошо видно, что значение *Acc* может оставаться относительно стабильным при существенном изменении *AUC* для соответствующего показателя. Например, в этой публикации для проточной цитофлуориметрии были получены значения *Acc* в диапазоне от 0,802 до 0,885 и соответствующие им значения *AUC* в диапазоне от 0,637 до 0,831. Если в качестве меры эффективности теста использовать *Acc*, то можно сказать, что тест обладает неплохой эффективностью (*Acc*>0,8) в то время, как на самом деле его скрининговая эффективность может быть недостаточной (*AUC*=0,637) или вполне приемлемой (*AUC*=0,831) для практического использования в зависимости от особенностей его проведения.

Таким образом, показатель *Acc* не может служить надежным показателем реальной эффективности (дескриптором валидности) теста. Если читателя интересует скрининговая эффективность маркера (то есть возможность на основе значения маркера отобрать бóльшую долю больных или здоровых субъектов), следует в первую очередь обращать внимание на интегральный показатель *AUC*, если интересует прогностическая эффективность (то есть возможность судить о вероятности наличия или отсутствия заболевания на основе значения маркера), то необходимо обращать внимание на *PBA*.

Отметим, что читателю необходимо четко осознавать, что кроется под термином *«accuracy»* (или *«diagnostic accuracy»*), т.к. иногда авторы [6, 7] понимают под этим термином *AUC* (в таком случае эта оценка имеет обычную для *AUC* интерпретацию). Очевидно, что *AUC* и *Acc* являются разными по своей сути оценками.

В случае, если читателю необходима некая общая мера сопряженности маркера и исхода, то лучше обратить свое внимание на *OR* и/или *RR*, а также их доверительные интервалы. Читателю необходимо помнить, что факт статистической значимости не означает, что маркер применим на практике (это необходимое, но не достаточное условие). *RR* является наиболее предпочтительным (с практической точки зрения) показателем сопряжённости, а *OR* – наиболее универсальным (это подробно рассмотрено в одной из предыдущих статей [1]). Тем не менее, при интерпретации этих показателей могут возникать трудности. Если положительное значение маркера сопряжено со снижением риска, то значение

$0<OR<RR<1$. Если маркер сопряжен с увеличением риска исхода (вероятно, более частая ситуация), то диапазон возможных значений лежит уже в значительно более широких пределах: $1<RR<OR<+\infty$. Напомним, что при интерпретации *OR* и *RR* читатель получает информацию не только о силе сопряженности маркера и исхода, но и направленности этой связи (уменьшение или увеличение риска исхода при наличии маркера).

Интуитивно более привлекательной является оценка, значения которой лежат в более определенных границах: от нуля (при отсутствии сопряженности маркера и исхода) до единицы при сильной сопряженности и увеличении риска при наличии маркера и до минус единицы, при сильной сопряженности и уменьшении риска при наличии маркера (по аналогии, например, с известным коэффициентом корреляции Пирсона). В связи с этим, когда возникает необходимость описать силу сопряженности маркера и исхода в дополнение к *OR* (или *RR*), целесообразно использовать иные оценки, например коэффициент корреляции Метьюса (*Matthews correlation coefficient – MCC*), предложенный в 1975 году [8]. Позже было предложено обобщение этой метрики для мультиклассового классификатора [9], то есть для тех случаев, когда номинальный признак имеет более двух категорий (например, не есть/нет заболевание, а заболевания нет, заболевание легкой/средней/тяжелой степени). *MCC* принимает значения от минус единицы до нуля, когда наличие маркера уменьшает риск исхода, и от нуля до единицы – когда увеличивает.

Этот показатель (в классическом его варианте для бинарного классификатора) вычисляется на основе всех четырех полей таблицы сопряженности 2×2 (*TP, FP, FN, TN*) и является одной из наиболее обобщенных и сбалансированных мер оценки эффективности теста. *MCC* также известен, как φ-коэффициент корреляции ($\varphi=\sqrt{(\chi^2/n)}$, где χ^2 – значение статистики хи-квадрат), а n – количество наблюдений. Ниже мы приведем формулу вычисления *MCC* не только с использованием *TP, FP, FN, TN,* но с других показателей (*Se*/*Sp* и др.). Видно, что этот показатель обобщает как скрининговую, так и прогностическую (диагностическую) эффективность маркера.

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}=$$

$$=\sqrt{(Se+Sp-1)\times(PPV+NPV-1)}=\sqrt{J_{Se\text{-}Sp}\times J_{PPV\text{-}NPV}}$$

MCC является одной их наиболее устойчивых при диспропорциональности матрицы оценок, в связи с чем она используется агентством по контролю за продуктами и лекарствами (FDA) США [10, 11] и также используется в нефрологии [12, 13], впрочем, надо отметить, что оценка *MCC* приводится чаще в контексте оценки результатов машинного обучения [13-16].

В первом примере (матрица слева) *MCC*=0,113 (при *Acc*=0,9), что свидетельствует о явно недостаточной для возможности практического применения маркера его сопряженности с исходом. Об этом же свидетельствуют и значения *AUC*=0,557 и *PBA*=0,557. В матрице справа *MCC*=0,402 (при *Acc*=0,63), что свидетельствует о более-менее приемлемой информативности маркера, главным образом, как скринингового (*AUC*=0,75), но не прогностического признака (*PBA*=0,662). Примерно то же самое можно сказать о примере 3, где *MCC*=0,268 (при *Acc*=0,862) и маркер обладает низкой эффективностью при скрининге (*AUC*=0,638) и как прогностический признак (*PBA*=0,63). В примере 4 *MCC*=0,468 (при *Acc*=0,939), а маркер обладает средней скрининговой (*AUC*=0,734) и прогностической эффективностью (*PBA*=0,734). Очевидно, что показатель *Acc* может вводить в заблуждение.

Кратко рассмотрим некоторые особенности применения оценки *MCC*, которые слабо представлены в имеющихся источниках. Значение *MCC* может существенно снижаться, когда $P_D>>P_M$ или $P_D<<P_M$ (рисунок 1Б), поскольку при этом неизбежно будет возрастать доля неверных классификаций: в первом случае – *FN* (т.к не все больные субъекты являются носителем маркера), во втором случае – *FP* (т.к. маркер будет нередко встречаться и у здоровых). Доля верных классификаций (*TP* и *TN*), соответственно, будет снижаться. Иными словами, малое значение *MCC* не обязательно означает, что маркер неэффективен. Важно другое. Максимально возможное значение *MCC* определяется *OR* (то есть силой сопряженности маркера и исхода) и связано с ним отношением:

$$MCC_{max}=\frac{\sqrt{OR-1}}{\sqrt{OR+1}}$$

В отличие от других диаграмм на рисунке 1, максимальное значение при заданном *OR* фиксировано. Таким образом, читатель может быть убежден, что при наличии статистической значимости и большом значении *MCC* тест будет эффективен.

Необходимо помнить, что показатель *MCC* может адекватно оценить силу сопряженности маркера и исхода в случае, если выборка была сформирована таким образом, что соотношения групп (с маркером/без маркера или больные/здоровые) не определялось непосредственно исследователем, а отражает истинную распространенность маркера и заболевания. То есть если исследование близко к популяционному: например, анализируется выборка пациентов с определенным заболеванием (целевая популяция) и оценивается частота исходов в зависимости от значения маркера. При когортном исследовании, когда выборка формируется по принципу «есть маркер/нет маркера», или при исследовании «случай-контроль», когда выборка формируется по принципу «есть исход/нет исхода», показатель *MCC* может давать искажённое представление о сопряженности маркера и исхода.

2. Обобщённая оценка скриниговой и прогностической информативности диагностического теста. До этого момента мы описали ряд оценок, которые характеризуют скрининговую (*Se, Sp, AUC*) или прогностическую (*PPV, NPV, PBA*) эффективность маркера. Кроме этого, мы рассмотрели основные общие меры сопряженности маркера и исхода (*OR, RR, Acc, MCC*). Интересной оценкой являются *F-меры*, и в частности – самая распространенная из них – *F1*. Показатель *F1* представляет собой сбалансированную обобщенную оценку (гармоническое среднее) чувствительности и прогностической ценности положительного результата, которые в ассоциации с *F-мерами* чаще всего именуются как полнота отклика – *«recall»* (*Se*) и прецизионность – *«precision»* (*PPV*). В данном контексте чувствительность можно охарактеризовать как долю пациентов с исходом, которую можно выявить на основе значения маркера («полноту»), а *PPV* – как «точность», уверенность в том, что идентифицированный как субъект с исходом действительно болен. Эта оценка позволяет получить сбалансированную характеристику информативности теста:

$$F1 = 2 \times \frac{PPV \times Se}{PPV + Se} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Мера *F1* характеризует эффективность теста с необычной стороны. В отличие от других интегральных показателей, *F*-оценка не учитывает истинно отрицательные результаты. С одной стороны, это может несколько снижать ее объективность, увеличивая предвзятость оценки [17, 18]. Поскольку данный тест не учитывает количество истинно отрицательных результатов, он характеризует только количество положительных «срабатываний»: наличие болезни при положительном значении маркера. С другой стороны, этот показатель позволяет более полно оценить именно способность теста распознавать пациентов с болезнью (но не отличать больных от здоровых!). Обобщение этих двух параметров (*Se* и *PPV*) дает весьма полезную информацию, которая не может быть получена при помощи других интегральных показателей.

Диапазон значений этого показателя лежит в диапазоне от нуля до единицы. *F*-мера стремится к нулю, если точность или полнота отклика стремится к нулю. Значение, близкое к единице, свидетельствует о высокой эффективности маркера. Однако, как и показатель *Acc*, *F*-мера может давать извращенную информацию в условиях сильно несбалансированной матрицы.

Данная мера чаще используется в машинном обучении при распознавании текстов и поиске информации, но тем не менее, безусловно, обладает некоторым потенциалом и для оценки эффективности бинарного теста в медицине. Мера *F1* нередко используется в публикациях, посвященных заболеваниям почек (вероятно, даже чаще, чем *MCC*). Например, при помощи *F1* авторы оценивают точность моделей, позволяющих предсказать развитие ХБП 5 стадии у пациентов с IgA-нефропатией [19, 20], развитие острого повреждения почек у пациентов с онкологическими заболеваниями [21] или у пациентов отделений интенсивной терапии [22].

Поскольку при расчете этого показателя используются и *Se* (которая может быть оценена при популяционном исследовании или исследовании случай-контроль), и *PPV* (которая может быть оценена при популяционном исследовании или когортном исследовании) для адекватной его оценки выборка должна быть сформирована путем включения пациентов без учета того, является ли субъект носителем маркера или нет, болен субъект или здоров (популяционное исследование).

Значение *F*-оценок убывает по мере уменьшения частоты встречаемости маркера и распространенности заболевания, особенно если $P_D >> P_M$ или $P_D << P_M$ (рисунок 1В). Это вполне закономерно: при $P_M \to 0$ и $P_D \to 0$ $TP \to 0$, то есть больных субъектов с маркером очень мало, проще говоря, тесту становится все «сложнее» выдавать правильные положительные классификации. При $P_D >> P_M$ $Se \to 0$ (маркер позволит выявить лишь небольшую долю больных), при $P_D << P_M$ $PPV \to 0$ (в этом случае снижается вероятность заболевания при наличии маркера, поскольку маркер встречается и у здоровых). Все это снижет эффективность теста и, соответственно, приводит к уменьшению значения *F*-меры. В обратной, очень маловероятной ситуации ($P_M \to 1$ и $P_D \to 1$) тест будет практически во всех случаях будет давать верные положительные классификации ($Se \to 1$, $PPV \to 1$), соответственно *F*-мера $\to 1$.

3. ROC-анализ. В случае, когда маркер представляет собой не номинальный бинарный признак (например, наличие или отсутствие отеков), а непрерывный количественный (например, концентрацию креатинина), бывает важно выявить порог, который позволит наиболее эффективно решать определенные задачи при помощи теста. В результате значения маркера больше определенного интерпретируются как положительные, а меньшие или равные – как отрицательные. Таким образом, тест сводится к снова к бинарной классификации. Эта задача решается при помощи *ROC*-анализа (*ROC-analysis – receiver operating characteristic analysis*).

Качество бинарной классификации оценивается путем вычисления долей верно идентифицированных больных (то есть *Se* или *TPR – true positive rate*) и неверно идентифицированных здоровых, (то есть 1 – *Sp* или *FPR – false positive rate*).

Рассмотрим пример информативности кардиопульмональной рециркуляции – *CPR* (*cardiopulmonary recirculation*) для выявления сердечной недостаточности (СН) с сохраненным сердечным выбросом. Кардиопульмональная рециркуляция представляет собой отношение объемной скорости кровотока по артериовенозной фистуле к минутному объему

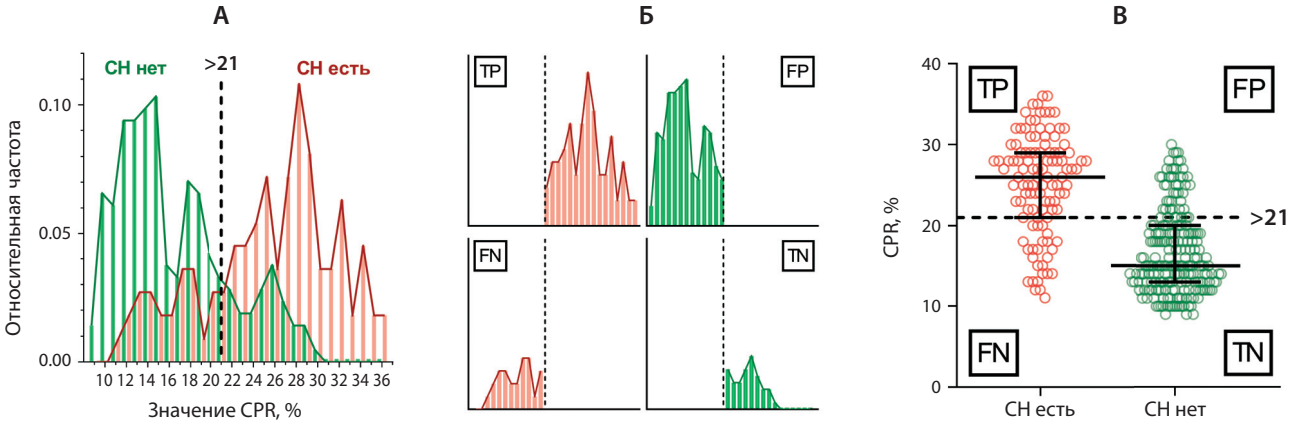

Рис. 2. Распределение показателя кардиопульмональной рециркуляции (CPR) у пациентов с (красный) и без (зеленый) сердечной недостаточности. *TP* – true positive (CPR > порогового, СН есть), *TN* – true negative (CPR≤порогового, СН нет), *FP* – false positive (CPR>порогового, СН нет), *FN* – false negative (CPR ≤ порогового, СН есть). На рисунке 2В также представлены медиана первый и третий квартили.

Fig. 2. Distribution of cardiopulmonary recirculation (CPR) in patients with (red) and without (green) heart failure (HF). *TP* – true positive (CPR > threshold, HF is present), *TN* – true negative (CPR ≤ threshold, HF is absent), *FP* – false positive (CPR > threshold, HF is absent), *FN* –false negative (CPR ≤ threshold, HF is present). Figure 2B also shows the median of the first and third quartiles.

кровообращения и отражает выраженность кардиотоксического действия артериовенозной фистулы [23].

На рисунке 2А представлено распределение этого признака (маркера) в группах пациентов с СН и без СН (исхода). Видно, что большие значения CPR чаще встречаются у пациентов с СН (красные столбики), меньшие – у пациентов без СН (зеленые столбики). Забегая вперед, отметим, что указанный параметр обладает достаточной информативностью – площадь под *ROC*-кривой (*AUC*) составляет 0,842 [95%CI 0,797; 0,88], $p<0,0001$. Пороговым значением, обеспечивающим наилучшую дискриминационную способность маркера, является >21% (то есть пациенты со значением CPR >21% могут быть отнесены к группе риска, поскольку у большей доли пациентов с СН CPR >21%, а у большей доли пациентов без СН CPR ≤21). Пока не будем спешить называть это пороговое значение оптимальным.

Для лучшего понимания *ROC*-анализа необходимо внимательно ознакомиться с таблицей 1. В рассматриваемом нами примере значения CPR лежат в интервале 9-36%. Если принять любое из этих значений в указанном диапазоне за пороговый уровень, то маркер сводится к бинарному (> или ≤ порогового значения). Таким образом, имея бинарный исход (есть/нет СН) мы может свести результаты классификации к уже знакомой нам четырехпольной таблице сопряженности два на два и подсчитать *TP* (CPR>порогового, СН есть), *TN* (CPR ≤ порогового, СН нет), *FP* (CPR > порогового, СН нет), *FN* (CPR ≤ порогового, СН есть) – рисунок 2Б (где пороговым значением выбрано CPR>21%). Кому-то, возможно, будет привычнее видеть распределите этого признака в группах, приведенное на рисунке 2В, где кроме уникальных значений приведены медианы, границы первого и третьего квартилей. Составление такой таблицы для каждого возможного значения маркера позволит вычислить ряд оценок, все из которых уже нам знакомы – таблица 1. А именно, в данном контексте нам будут интересны P_M, P_D, *Se*, *Sp*, *AUC*, $J_{Se\text{-}Sp}$.

Например, если принять пороговым значением CPR=9%, то все пациенты с СН будут иметь значение больше (*TP*=111, *FN*=0), только трое пациентов без СН будут иметь значение CPR ≤9 (*TN*) и 210 пациентов без СН будут иметь значение CPR >9% (*FP*). Такие же вычисления проведем для остальных пороговых значений CPR и вычислим необходимые оценки.

Из таблицы 1 видно, что мере увеличение порогового уровня CPR меняется P_M, поскольку та доля лиц, у которых мы может выявить наличие маркера, меняется: например, при значении CPR >11% маркер (положительное значение признака, то есть более порогового значения) можно выявить у 293 субъектов (*TP* + *FP*) из 324 (90,4%), в то время как если выбрать пороговым значением CPR >28%, то маркер можно выявить только у 39 субъектов (12%). При этом показатель P_D остается стабильным, что закономерно, т.к. количество субъектов с СН не меняется, меняется лишь наше представление о том, кого можно считать субъектом с СН, а кого – без СН на основе значения маркера – определенного порогового значения CPR.

По мере увеличения порогового значения CPR убывает количество *TP* и возрастает количество *TN*, поскольку все меньше субъектов с СН и все больше субъектов без СН будут носителями признака (значения CPR > порогового значения). Одновременно

Таблица 1 | **Table 1**

Зависимость показателей информативности маркера от порогового уровня кардиопульмональной рециркуляции (CPR). Заболеванием считали наличие сердечной недостаточности с сохранённым сердечным выбросом, а положительным значением маркера в каждой строчке – значение CPR больше, указного в первом столбце. Зеленая полупрозрачная заливка отражает долю от максимального значения по столбцу.

The dependence of the marker informativeness estimates on the threshold level of cardiopulmonary recirculation (CPR). The disease was considered the presence of heart failure with a preserved cardiac output, and a positive marker value in each row – the CPR value is greater than indicated in the first column. The green semi-transparent fill reflects the percentage of the maximum value for the column.

CPR, %	*TP*	*TN*	*FP*	*FN*	P_M	P_D	*Se*	*Sp*	*AUC*	$J_{Se\text{-}Sp}$
9	111	3	210	0	0,991	0,343	1	0,014	0,507	0,014
10	111	17	196	0	0,948	0,343	1	0,08	0,54	0,080
11	110	30	183	1	0,904	0,343	0,991	0,141	0,566	0,132
12	108	50	163	3	0,836	0,343	0,973	0,235	0,604	0,208
13	105	70	143	6	0,765	0,343	0,946	0,329	0,637	0,275
14	102	91	122	9	0,691	0,343	0,919	0,427	0,673	0,346
15	100	113	100	11	0,617	0,343	0,901	0,531	0,716	0,431
16	98	121	92	13	0,586	0,343	0,883	0,568	0,725	0,451
17	94	128	85	17	0,552	0,343	0,847	0,601	0,724	0,448
18	90	143	70	21	0,494	0,343	0,811	0,671	0,741	0,482
19	89	157	56	22	0,448	0,343	0,802	0,737	0,769	0,539
20	86	166	47	25	0,41	0,343	0,775	0,779	0,777	0,554
21	83	173	40	28	0,38	0,343	0,748	0,812	0,78	0,56
22	78	179	34	33	0,346	0,343	0,703	0,84	0,772	0,543
23	73	183	30	38	0,318	0,343	0,658	0,859	0,758	0,517
24	67	187	26	44	0,287	0,343	0,604	0,878	0,741	0,482
25	59	193	20	52	0,244	0,343	0,532	0,906	0,719	0,438
26	55	201	12	56	0,207	0,343	0,495	0,944	0,72	0,439
27	47	206	7	64	0,167	0,343	0,423	0,967	0,695	0,391
28	35	209	4	76	0,12	0,343	0,315	0,981	0,648	0,297
29	26	212	1	85	0,083	0,343	0,234	0,995	0,615	0,230
30	22	213	0	89	0,068	0,343	0,198	1	0,599	0,198
31	18	213	0	93	0,056	0,343	0,162	1	0,581	0,162
32	11	213	0	100	0,034	0,343	0,099	1	0,55	0,099
33	9	213	0	102	0,028	0,343	0,081	1	0,541	0,081
34	4	213	0	107	0,012	0,343	0,036	1	0,518	0,036
35	2	213	0	109	0,006	0,343	0,018	1	0,509	0,018
36	0	213	0	111	0	0,343	0	1	0,5	0,000

(но не вполне симметрично) убывает и значение *FP*, но и возрастает значение *FN*, поскольку реже маркер будет обнаруживаться у здоровых субъектов со значением CPR ≤ порогового значения (*FP*) и все чаще маркера не будет у субъектов без СН (*FN*), то есть значение CPR у субъектов без СН будет ниже порогового.

Графическим выражением результатов *ROC*-анализа является *ROC*-кривая (и ее 95% доверительный интервал, 95%CI – 95% confidence interval) – рисунок 3А. Суть этого графика сводится к следующему. Для каждого из возможных значений признака вычисляются доли *Se* и *Sp*. Каждая точка на графике – значения *Se* и *FPR* (1 – *Sp*) для одного из уникальных значений количественного признака (в данном случае – CPR). Далее через все эти точки последовательно проводится линия.

В рутинной практической работе доктор, как правило, сталкивается к решению задач бинарной классификации (есть или нет заболевание, назначать или не назначать специфическое лечение) В этой ситуации важно определить оптимальное пороговое значение признака (в данном случае – CPR), которое позволит однозначно классифицировать субъект, как имеющий или не имеющий заболевание. Для решения этой задачи существует как минимум два подхода.

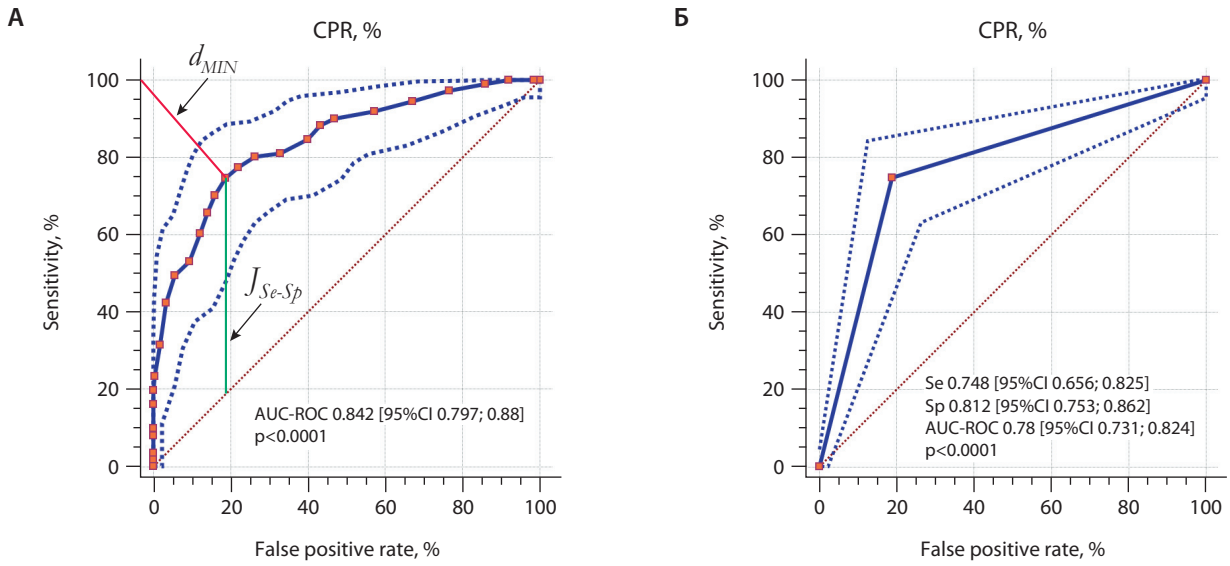

Рис. 3. ROC-кривая, описывающая дискриминационную способность маркера (кардиопульмональной рециркуляции).

Fig. 3. ROC-curve describing the discriminative ability of the marker (cardiopulmonary recirculation).

Первый подход – утилитарный. В данном случае выбирается пороговое значение маркера, обеспечивающее наилучшую его дискриминационную способность. Напомним, что *Se* и *Sp* характеризуют скрининговую (но не прогностическую!) эффективность маркера, то есть способность выявлять субъектов с исходом и без исхода (однако эти показатели ничего не скажут о вероятности заболевания при положительном значении маркера и вероятности отсутствия исхода при отрицательном значении маркера). Чем лучше маркер позволяет выявлять здоровых и больных (в данном примере – с СН и без СН), тем он эффективнее. Соответственно, исследователя при таком подходе интересует пороговое значение признака, обеспечивающее наибольшие значение *Se*, *Sp* и *AUC* ($AUC=(Se+Sp)/2$). Вернемся к таблице 1, из которой видно, что по мере возрастания *Sp* убывает *Se*. Это закономерно, поскольку чем больше пороговое значение CPR, тем больше субъектов со значением маркера ≤ порогового могут быть отнесены к субъектам без СН на основе значения маркера при скрининге (*Sp*) и тем меньше субъектов с СН имеют значение маркера больше порогового (*Se*). Например, если выбрать пороговым значением CPR >11%, то это позволит выявить 99,1% субъектов с СН, но только 14,1% субъектов без СН. Если пороговым значением CPR выбрать значение CPR ≤28%, то это позволить выявить только 31,5% с СН, но 98,1% субъектов без СН.

Выбрать оптимальное пороговое значение в данном случае можно на основе максимального значения показателя *AUC* или индекса Юдена ($J_{Se-Sp}=Se+Sp-1$), которые были подробно рассмотрены нами ранее [2]. На основе этих оценок мы получим оптимальное пороговое значение CPR >21%, это обеспечит *Se*=0,748 и *Sp*=0,812, *AUC*=0,78, J_{Se-Sp}=0,56. На рисунке 3А приведена *ROC*-кривая, отражающая связь показателя CPR и СН, а также на ней указан показатель J_{Se-Sp}=0,56 (данный индекс в своем графическом выражении представляет собой расстояние от точки на *ROC*-кривой до диагональной изолинии, равноудаленной от осей), а также весьма популярный способ найти оптимальное пороговое значение – определить минимальное расстояние до левого верхнего угла (d_{MIN}, который также известен, как *K*-индекс [24]), то есть до точки с *Se*=1 и *Sp*=1 (напомним, что по оси абсцисс на графике *ROC*-кривой отложены значения FPR, то есть 1–*Sp*). Если учесть, что координаты точек на *ROC*-кривой – это *Se* и 1–*Sp*, то найти длину отрезка d_{MIN} для каждой точки можно очень просто, используя теорему Пифагора. В данном случае

$$d_{MIN}=\sqrt{(1-Se)^2+(1-Sp)^2}=$$
$$=\sqrt{(1-0{,}748)^2+(1-0{,}812)^2}=0{,}314.$$

При выборе оптимального порогового значения можно исходить из вероятностных или клинических соображений. При проведении традиционного *ROC*-анализа ложноположительные и ложноотрицательные результаты имеют равный вес (важность с клинической, экономической или любой иной точки зрения), что позволяет использовать координаты точек для определения порогового значения [25]. Несмотря на то, что d_{MIN} и индекс Юдена в большинстве случаев дают одно и то же пороговое значение, это происходит не всегда. В одной из работ [26] приводятся аргументы в пользу преимущественного использования индекса Юдена, основным из кото-

рых является тот факт, что этот показатель в отличие от d_{MIN} имеет простую клиническую интерпретацию. А именно: индекс Юдена представляет собой разность между долей верно идентифицированных больных (*Se*) и долей неверно идентифицированных здоровых (1 – *Sp*): $Se-(1-Sp)=Se+Sp-1$. Иными словами, этот показатель максимизирует разницу между истинно положительными и ложноположительными результатами теста: чем больше первых и чем меньше вторых, тем тест лучше. В свою очередь К-индекс является абстрактной мерой и не имеет вероятностной интерпретации.

Значение *AUC* на рисунке 3А (0,842) не соответствует значению *ROC-AUC,* полученном нами в таблице 1 (и на рисунке 3Б) для порогового значения CPR >21% (0,78). Ничего удивительного в этом нет, поскольку в первом случае указана площадь под всей *ROC*-кривой (рисунок 3А), а во-втором – только для одной точки. Во втором случае *ROC*-кривая принимает кусочно-линейный вид (рисунок 3Б). Читатель всегда должен понимать, какое значение *AUC* приводится. Когда идет речь об информативности маркера как таковой – вероятнее всего приводится площадь под всей *ROC*-кривой (ограниченной всеми точками, с соответствующими координатами всех уникальных значений признака). Если указывается одно пороговое значение, вероятнее всего, авторы публикации имеют в виду обобщённую оценку *Se* и *Sp* для конкретного порогового значения признака.

Второй подход – адаптация маркера под конкретные задачи. В случае, когда пороговым значением маркера CPR выбрано 21%, маркер обладает наибольшей дискриминантной мощностью, то есть способностью выявлять пациентов с СН и без СН при скрининге. То есть достигается оптимальный с утилитарной точки зрения баланс между чувствительностью и специфичностью. Однако такой компромисс не всегда является оптимальным решением.

Приоритет может быть отдан высокой чувствительности (частично пожертвовав специфичностью), если важно «упустить» минимальное количество больных, даже при условии «гипердиагностики». Допустим, важно выявить не менее 90% пациентов с СН. Для этого пороговым значением CPR может быть выбрано меньшее значение, например, CPR >15%. Это обеспечит *Se*=0,901, то есть позволит выявить 90,1% пациентов с СН, поскольку у такой доли субъектов с СН значение CPR больше порогового. Однако в этом случае тест обладает меньшей специфичностью (*Sp*=0,531), поскольку у меньшей доли субъектов без СН значение CPR будет меньше или равно пороговому. Иными словами, если маркер предназначен для конкретных целей (что бывает нередко), это существенным образом меняет подход к выбору порогового значения. Например, в данном случае (CPR >15%) мы сможем выявить подавляющее большинство пациентов с СН (более 90%). Однако к этой группе будут отнесены и 46,9% (1 – *Sp*=0,469, то есть 46,9%) от всех пациентов без СН (в очередной раз обратим ваше внимание на то, что Se не является вероятностью наличия СН при значении CPR >15%, а *Sp* – вероятностью отсутствия СН при CPR ≤15%).

В противоположном случае предпочтение может быть отдано специфичности за счет чувствительности. Если выбрать пороговым значение, например, CPR >25, то это обеспечит *Se*=0,532 при *Sp*=0,906. То есть, такой более высокий порог позволит выявить всего 53,2% пациентов с СН, но 90,6% пациентов без СН. Примером может служить профотбор. Такой подход применим, когда тест должен быть высокоспецифичным, если важно отобрать большую долю субъектов без исхода интереса, не допустить «гипердиагностики», даже при условии, что могут быть пропущены некоторые пациенты с заболеванием.

Такой подход применим, если он обеспечивает приемлемое значение интегральных показателей информативности маркера, среди которых наиболее часто используется *AUC*. В первом случае *AUC*=0,716, а во втором – *AUC*=0,719, что свидетельствует о том, что такой маркер будет обладать средней скрининговой эффективностью. Для высокоэффективного маркера *AUC* должна составлять 0,8 и более.

Изменяя пороговое значение CPR, мы меняем доли субъектов с СН и без СН, которые могут быть выявлены при скрининге. Это просто понять, если посмотреть на рисунки 2А и В: мысленно перемещая пунктирную линию по оси абсцисс на рисунке 2А или по оси ординат на рисунке 2В, мы меняем пороговое значение CPR и, соответственно, доли пациентов, которые будут отнесены к больным и здоровым на основе значения этого показателя. При увеличении порогового значения CPR мы увеличиваем долю субъектов без СН, которые могут быть выявлены на основе значения CPR, поскольку увеличивается доля субъектов без СН со значением CPR меньшим или равным пороговому (то есть, увеличиваем специфичность). Однако при этом уменьшается доля субъектов с СН, которые могут быть выявлены при скрининге, поскольку уменьшается доля субъектов с СН со значением больше порогового (то есть, уменьшается специфичность). И наоборот, если пороговое значение CPR уменьшается, то это сопровождается увеличением выявленной на основе значения этого маркера при скрининге доли субъектов с СН (увеличивается специфичность) и снижается доля верно идентифицированных субъектов без СН (уменьшается специфичность).

Несколько отклоняясь от нити повествования, считаем важным отметить, что, возможно, интуитивно ожидаемо, что для более эффективного выявления субъектов с СН необходимо выбрать большее значение CPR. Однако это не так. Выбор большего порогового значения признака будет приводить к уменьшению доли субъектов с бóльшим его

Таблица 2 | **Table 2**

Информативность концентрации NGAL в плазме крови при прогнозировании различных вариантов ОПП. Приведены пороговые значения, соответствующие 95% чувствительности, максимальному значению индекса Юдена и 95% специфичности. Se – чувствительность, Sp – специфичность, PPV – прогностическая ценность положительного результата теста, NPV – прогностическая ценность отрицательного результата теста. Цитируется по [27] с изменениями.

Informative value of NGAL blood plasma concentration in predicting various AKI variants. The threshold values corresponding to 95% sensitivity, the maximum value of the Yuden index, and 95% specificity are given. Se – sensitivity, Sp – specificity, PPV – predictive value of a positive test result, NPV – predictive value of a negative test result. Quoted from [27] with changes.

Конечная точка	Критерий	Пороговый уровень, нг/мл	*Se*	*Sp*	*PPV*	*NPV*
ОПП без ЗПТ	*Se*=0,95	71	0,95	0,22	0,237	0,945
	J-индекс	165	0,66	0,73	0,384	0,894
	Sp=0,95	311	0,3	0,95	0,604	0,842
ОПП с ЗПТ	*Se*=0,95	162	0,95	0,59	0,135	0,994
	J-индекс	214	0,87	0,71	0,168	0,988
	Sp=0,95	546	0,26	0,95	0,259	0,95

значением у пациентов с исходом (см. рисунок 1В). Это позволит лишь существенно повысить вероятность исхода при положительном значении маркера (то есть, приведет к увеличению показателя *PPV*). Если исследователь, напротив, выберет меньшее пороговое значение изучаемого признака, то это, как правило, приводит к увеличению вероятности отсутствия исхода при отрицательном значении маркера (то есть *NPV*), но доля верно идентифицированных лиц без исхода (*Sp*) будет меньше.

Определение оптимального порогового значения маркера является концептуальной основой для достижения компромисса между чувствительностью и специфичностью теста. Далеко не всегда утилитарный выбор этого значения в полной мере отвечает целям применения теста.

В таблице 2 приведена часть данных из представленных в недавно опубликованном мета-анализе [27], где оценена информативность ассоциированного с желатиназой нейтрофилов липокалина (neutrophil gelatinase-associated lipocalin – NGAL) при прогнозировании различных вариантов тяжести острого повреждения почек (ОПП). Хорошо заметно, как сильно различается пороговые значения концентрации NGAL в плазме крови, соответствующие 95% чувствительности, максимальному значению индекса Юдена и 95% специфичности. По мере увеличения порогового значения снижается *Se* и возрастает *Sp* и *PPV*. При этом видно, что, например, пороговое значение (165 нг/мл) для прогнозирования ОПП, не требующего ЗПТ, выбранное на основе максимального значения индекса Юдена, обеспечивает значения *Se* и *Sp* (как и *PPV*), явно недостаточные для практического использования этого маркера. В то же время, если нужно обеспечить большое значение чувствительности (*Se*=0,95), то есть максимальную эффективность для выявления (но не подтверждения!) лиц с повышенным риском ОПП целесообразно выбрать меньшее пороговое значение (71 нг/мл). Если необходимо обеспечить, например, большую вероятность развития ОПП при положительном значении маркера (*PPV*), целесообразно выбрать большее пороговое значение. Существуют также и другие работы, где приводится несколько пороговых значений количественных показателей, связанных с заболеваниями почек, для конкретных фиксированных *Se* и *Sp* [28-31].

В большинстве работ, описывающих эффективность маркеров, для определения оптимального порогового значения используется *K*-индекс (d_{MIN}) или *J*-индекс Юдена. Это оправдано, когда маркер обладает большой информативностью и очень сильно сопряжен с исходом. Как правило, в таких случаях *AUC* приближается или превышает 0,9, так же, как и значения *Se* и *Sp*. В случае менее эффективного маркера оптимальное пороговое значение («*cut-off*») количественного признака может быть выбрано не только на основе максимального значения индекса Юдена, *AUC* или минимального значения индекса d_{MIN}, но и на основе других оценок, позволяющих адаптировать маркер для конкретных практических задач. Указание на это содержится не только в публикациях в медицинских журналах [25, 32] и руководствах по статистике [33], адресованных широкому кругу читателей, но и, например, в главе, посвященной особенностям статистического анализа в книге «*Biomarkers of Kidney Disease*» [34]. В современных статистических программах, в частности – в R, представлено несколько библиотек [35, 36], которые позволяют определить оптимальное пороговое значение на основе максимизации *PPV* и *NPV*, *Se* и *Sp*, суммы этих оценок (или минимизации разности между ними), максимизации *OR*, *RR*, отношения правдоподобия, минимизации *p*-value или максимизации статистики χ^2 и другими способами.

Заключение

Таким образом, читателю необходимо помнить, что оценка информативности диагностических признаков не так проста, как может показаться с первого взгляда.

1. В случае, если в публикации в качестве общей оценки эффективности теста приводится показатель *Acc*, читателю нужно быть осторожным в случае редкого заболевания (малой кумулятивной частоты исходов) и в первую очередь обратить внимание на *AUC* и/или *PBA*. Большие значения (0,8 и более) именно этих показателей свидетельствуют о хорошей дискриминационной способности диагностического теста (скрининговой или прогностической соответственно). Изолированная оценка Acc не позволяет сделать такого вывода. Более приемлемой альтернативой можно *Acc* считать *MCC*, которая является обобщенной оценкой эффективности теста.

2. *F-меры* позволяют получить обобщенную оценку скрининговой и прогностической информативности диагностического теста, но не учитывает количество истинно отрицательных результатов. В связи с этим, нужно с осторожностью интерпретировать этот показатель. Он не дает информации о том, насколько хорошо тест работает вообще. Иными словами, если несколько пожертвовать точностью в пользу доступности: не позволяет судить о том, насколько эффективно тест позволяет отличать больных от здоровых, а позволяет судить лишь о его способности выявлять больных.

3. Далеко не всегда оптимальное пороговое значение количественного показателя определяется в результате утилитарной максимизации оценок чувствительности и специфичности. Как правило, так поступают, когда маркер очень сильно сопряжен с исходом, а, например *Se* и *Sp* имеют значения 0,9 и больше. На практике чаще выбирают оптимальное пороговое значение на основании иных принципов, адаптируя маркер к определенным условиям и тем самым повышая эффективность его применения для какой-то определенной цели (самый простой пример: для проведения скрининга или же – используя его как прогностический признак).

4. Читателю необходимо критически оценивать публикации и помнить о том, что дизайн исследования определяет спектр оценок, которые можно получить в результате исследования. Скрининговую эффективность маркера (*Se, Sp, AUC*) можно оценить при проведении исследования «случай-контроль», прогностическую эффективность – при проведении когортного исследования. В случае беспристрастного формирования выборки из целевой популяции в ходе популяционного исследования (поперечное исследование) можно получить любые оценки сопряженности маркера и заболевания, а также – информативности диагностического теста.

Авторы заявляют об отсутствии конфликта интересов

The authors declare no conflict of interest

Работы были выполнены с использованием средств гранта Президента Российской Федерации для государственной поддержки молодых российских ученых № МК-63.2020.7.

Funding: grant of the President of the Russian Federation for state support of young Russian scientists No. MK-63.2020.7.

Персональный вклад авторов в рукопись:

Зулькарнаев А.Б. – общая концепция работы, статистический анализ, подготовка иллюстраций и таблиц, написание текста работы, окончательное редактирование текста рукописи.

Паршина Е.В. – работа с литературными источниками, поиск примеров использования различных оценок в публикациях, окончательное редактирование текста рукописи, консультирование членов авторского коллектива по различным аспектам заболеваний почек, подготовка абстракта и перевод его на английский язык.

Федулкина В.А. – работа с литературными источниками, оформление работы, окончательное редактирование текста рукописи.

Author contribution:

Zulkarnaev A.B. – general concept of the manuscript, statistical analysis, preparation of tables and figures, writing the draft of the manuscript and its final editing.

Parshina E.V. – literature searching, search for relevant examples of various assessments usage, final editing of the manuscript, advising on various aspects of kidney disease, abstract preparation and translation into English.

Fedulkina V.A. – literature searching, design of the manuscript, final editing of the manuscript.

Список литературы

1. *Зулькарнаев А.Б.* «Подводные камни» статистического анализа и клинической интерпретации полученных оценок на примере пациентов с хронической болезнью почек. Часть I: оценка риска. Нефрология и диализ. 2019; 21(4): 419-429.

Zulkarnaev A.B. Pitfalls of statistical analysis and clinical interpretation of the estimates in patients with chronic kidney disease. Part I: risk assessment. Nephrology and dialysis. 2019; 21(4): 419-429. doi: 10.28996/2618-9801-2019-4-419-429

2. *Зулькарнаев А.Б. Паршина Е.В.* «Подводные камни» статистического анализа и клинической интерпретации полученных оценок на примере пациентов с хронической болезнью почек. Часть III: Оценка информативности биомаркеров. Нефрология и диализ. 2021; 23(1): 105-118.

Zulkarnaev A.B., Parshina E.V. Pitfalls of statistical analysis and clinical interpretation of the estimates on the example of patients with chronic kidney disease Part III: Evaluating the informativeness of biomarkers. Nephrology and dialysis. 2021; 23(1): 105-118. doi: 10.28996/2618-9801-2021-1-105-118

3. *Koren W., Koldanov R., Pronin V.S. et al.* Amiloride-sensitive Na+/H+ exchange in erythrocytes of patients with NIDDM: a prospective study. Diabetologia. 1997; 40(3): 302-6. doi: 10.1007/s001250050678.

4. *Hänninen E.L., Denecke T., Stelter L. et al.* Preoperative evaluation of living kidney donors using multirow detector computed tomography: comparison with digital subtraction angiography and intraoperative findings. Transpl Int. 2005; 18(10):1134-41. doi: 10.1111/j.1432-2277.2005.00196.x.

5. *Peräsaari J.P., Jaatinen T., Merenmies J.* Donor-specific HLA antibodies in predicting crossmatch outcome: Comparison of three different laboratory techniques. Transpl Immunol. 2018; 46: 23-28. doi: 10.1016/j.trim.2017.11.002.

6. *Nixon A.C., Bampouras T.M., Pendleton N. et al.* Diagnostic Accuracy of Frailty Screening Methods in Advanced Chronic Kidney Disease. Nephron. 2019; 141(3): 147-155. doi: 10.1159/000494223.

7. *Kovesdy C.P., Molnar M.Z., Czira M.E. et al.* Diagnostic accuracy of serum parathyroid hormone levels in kidney transplant recipients with moderate-to-advanced CKD. Nephron Clin Pract. 2011; 118(2): c78-85. doi: 10.1159/000320318.

8. *Matthews B.W.* Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975; 405(2): 442-451. doi:10.1016/0005-2795(75)90109-9

9. *Gorodkin J.* Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem. 2004; 28(5-6):367-74. doi: 10.1016/j.compbiolchem.2004.09.006.

10. *Shi L., Campbell G., Jones W.D. et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010; 28(8): 827-38. doi: 10.1038/nbt.1665.

11. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol. 2014; 32(9): 903-14. doi: 10.1038/nbt.2957.

12. *Yin W.J., Yi Y.H., Guan X.F. et al.* Preprocedural Prediction Model for Contrast-Induced Nephropathy Patients. J Am Heart Assoc. 2017; 6(2):e004498. doi: 10.1161/JAHA.116.004498.

13. *Singh N.P., Bapi R.S., Vinod P.K.* Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma. Comput Biol Med. 2018; 100: 92-99. doi: 10.1016/j.compbiomed.2018.06.030.

14. *Kannan S., Morgan L.A., Liang B. et al.* Segmentation of Glomeruli Within Trichrome Images Using Deep Learning. Kidney Int Rep. 2019; 4(7): 955-962. doi: 10.1016/j.ekir.2019.04.008.

15. *Hu L., Li H., Cai Z. et al.* A new machine-learning method to prognosticate paraquat poisoned patients by combining coagulation, liver, and kidney indices. PLoS One. 2017; 12(10):e0186427. doi: 10.1371/journal.pone.0186427.

16. *Kocak B., Yardimci A.H., Bektas C.T. et al.* Textural differences between renal cell carcinoma subtypes: Machine learning-based quantitative computed tomography texture analysis with independent external validation. Eur J Radiol. 2018; 107: 149-157. doi: 10.1016/j.ejrad.2018.08.014.

17. *Sokolova M., Lapalme G.* A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009; 45(4): 427-437. doi: 10.1016/j.ipm.2009.03.002

18. *Powers D.M.W.* Evaluation: from precision, recall and F-factor to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies. 2011; 2: 37-63.

19. *Diciolla M., Binetti G., Di Noia T. et al.* Patient classification and outcome prediction in IgA nephropathy. Comput Biol Med. 2015; 66: 278-286. doi:10.1016/j.compbiomed.2015.09.003

20. *Liu Y., Zhang Y., Liu D. et al.* Prediction of ESRD in IgA Nephropathy Patients from an Asian Cohort: A Random Forest Model. Kidney Blood Press Res. 2018; 43(6): 1852-1864. doi: 10.1159/000495818.

21. *Park N., Kang E., Park M. et al.* Predicting acute kidney injury in cancer patients using heterogeneous and irregular data. PLoS One. 2018; 13(7):e0199839. doi: 10.1371/journal.pone.0199839.

22. *Morid M.A., Sheng O.R.L., Del Fiol G. et al.* Temporal Pattern Detection to Predict Adverse Events in Critical Care: Case Study With Acute Kidney Injury. JMIR Med Inform. 2020; 8(3):e14272. doi: 10.2196/14272.

23. *Lok C.E., Huber T.S., Lee T. et al.* KDOQI Clinical Practice Guideline for Vascular Access: 2019 Update. Am J Kidney Dis. 2020; 75(4 Suppl 2):S1–S164. doi: 10.1053/j.ajkd.2019.12.001.

24. *Kallner A.* Laboratory Statistics. Methods in Chemistry and Health Sciences. 2nd Edition. Elsevier. 2018. 174 p.

25. *Tripepi G., Jager K.J., Dekker F.W., Zoccali C.* Diagnostic methods 2: receiver operating characteristic (ROC) curves. Kidney Int. 2009; 76(3): 252-6. doi: 10.1038/ki.2009.171.

26. *Perkins N.J., Schisterman E.F.* The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol. 2006; 163(7): 670-5. doi: 10.1093/aje/kwj063.

27. *Albert C., Zapf A., Haase M. et al.* Neutrophil Gelatinase-Associated Lipocalin Measured on Clinical Laboratory Platforms for the Prediction of Acute Kidney Injury and the Associated Need for Dialysis Therapy: A Systematic Review and Meta-analysis. Am J Kidney Dis. 2020; 76(6): 826-841.e1. doi: 10.1053/j.ajkd.2020.05.015.

28. *Couchoud C., Pozet N., Labeeuw M., Pouteil-Noble C.* Screening early renal failure: cut-off values for serum creatinine as an indicator of renal impairment. Kidney Int. 1999; 55(5): 1878-84. doi: 10.1046/j.1523-1755.1999.00411.x.

29. *Twerenbold R., Wildi K., Jaeger C. et al.* Optimal Cutoff Levels of More Sensitive Cardiac Troponin Assays for the Early Diagnosis of Myocardial Infarction in Patients With Renal Dysfunction. Circulation. 2015; 131(23): 2041-50. doi: 10.1161/CIRCULATIONAHA.114.014245.

30. *Candela-Toha Á., Pardo M.C., Pérez T. et al.* Estimated glomerular filtration rate is an early biomarker of cardiac surgery-associated acute kidney injury. Nefrologia. 2018; 38(6): 596-605. English, Spanish. doi: 10.1016/j.nefro.2018.01.002.

31. *Waikar S.S., Betensky R.A., Emerson S.C., Bonventre J.V.* Imperfect gold standards for kidney injury bio-

marker evaluation. J Am Soc Nephrol. 2012; 23(1): 13-21. doi: 10.1681/ASN.2010111124.

32. *Ray P., Le Manach Y., Riou B., Houle T.T.* Statistical evaluation of a biomarker. Anesthesiology. 2010; 112(4): 1023-40. doi: 10.1097/ALN.0b013e3181d47604.

33. *Hoffman J.* Biostatistics for Medical and Biomedical Practitioners. 2nd Edition. Academic Press. 2019. 734 p.

34. *Edelstein C.* Biomarkers of Kidney Disease. 2nd Edition. Academic Press. 2016. 632 p.

35. *Thiele C., Hirschfeld G.* cutpointr: Improved Estimation and Validation of Optimal Cutpoints in R. arXiv [stat.CO]. 2020. Available from: http://arxiv.org/abs/2002.09209.

36. *López-Ratón M., Rodríguez-Álvarez M.X., Cadarso-Suárez C., Gude F.* OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of Statistical Software. 2014; 61(8): 1-36. doi: 10.18637/jss.v061.i08

Дата получения статьи: 04.11.2021

Дата принятия к печати: 06.02.2022

Submitted: 04.11.2021

Accepted: 06.02.2022