© А. В. Рубанович[1],
Н. Н. Хромов-Борисов[2-4]

[1] ФГБУН «Институт общей генетики им. Н. И. Вавилова РАН», Москва;

[2] ГБОУ ВПО «Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова Минздрава РФ»;

[3] ФГУ «Российский научно-исследовательский институт гематологии и трансфузиологии Федерального медико-биологического агентства», Санкт-Петербург;

[4] ФГБУ «Российский научно-исследовательский институт травматологии и ортопедии им. Р. Р. Вредена Минздрава РФ», Санкт-Петербург

❀ **Представлен свод формул для показателей распознающей и предсказательной способностей бинарных генетических тестов. Охарактеризована их зависимость от распространенности заболевания и частоты встречаемости генетического маркера. Показано, что при отношении шансов $OR < 2{,}2$ маркер обладает заведомо низкой прогностической эффективностью во всех смыслах при любых частотах заболевания и маркера. Маркер может быть хорошим классификатором, если $OR > 5{,}4$, но лишь при условии, что его популяционная частота достаточно высока ($p_M > 0{,}3$). Приведены формулы, позволяющие в исследованиях типа «случаи–контроли» получать косвенные оценки абсолютных и относительных рисков носительства маркера.**

❀ **Ключевые слова:** генетические ассоциативные исследования; отношение шансов; AUC; предсказательный генетический тест.

Поступила в редакцию 03.08.2012
Принята к публикации 10.01.2012

УДК 57.087.1

# ТЕОРЕТИЧЕСКИЙ АНАЛИЗ ПОКАЗАТЕЛЕЙ ПРЕДСКАЗАТЕЛЬНОЙ ЭФФЕКТИВНОСТИ БИНАРНЫХ ГЕНЕТИЧЕСКИХ ТЕСТОВ

## *ВВЕДЕНИЕ*

Повсеместное распространение исследований статистических связей между генотипом и предрасположенностью к широко распространенным заболеваниям породило острую дискуссию о методах оценки прогностической эффективности маркеров, выявляемых в результате этих работ (Poste, 2011; Kraemer et al., 2011; Pepe et al., 2010; Kraft et al., 2009; Jakobsdottir et al., 2009; Tan et al., 2004). В большинстве случаев авторы сходятся во мнении о том, что высокие значения показателей сопряженности маркера с признаками не гарантируют возможности использования этого маркера для прогноза фенотипического проявления генотипа. В частности, многие авторы подчеркивают, что статистически высоко значимая сопряженность заболевания с генетическим маркером является необходимым, но не достаточным условием возможности использовать такой маркер для предсказания предрасположенности к заболеванию. Так, например, многочисленные гены, выявляемые при широкогеномном сканировании как сопряженные с раком предстательной железы, лишь на несколько процентов увеличивают предсказательную эффективность традиционных биомаркеров (PSA, Gleason score) (см., например, Aly et al., 2011 и редакторский комментарий Bjartell, 2011).

В этой связи нами было предпринято теоретическое исследование ситуаций, возникающих при попытках описания статистических связей «генотип — бинарный признак». В первую очередь нас интересовал вопрос: какие значения *OR* (отношения шансов) могут обнадежить исследователя? При каких *OR* на основе выявленной генетической ассоциации может быть создан эффективный биомаркер предрасположенности к заболеванию? Мы покажем, что ответы на эти вопросы существенно зависят от распространенности заболевания и частоты встречаемости маркера. Цель публикации — всесторонне исследовать функциональную зависимость стандартных показателей эффективности теста от трех независимых параметров: *OR,* популяционная частота маркера ($p_M$) и распространенность заболевания ($p_D$).

## *ОБЗОР ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ БИНАРНОГО ТЕСТА*

Количество предложенных мер сопряженности качественных признаков давно превышает все разумные пределы. В работе (Tan et al., 2004) перечислен 21 показатель ассоциирования, характеризующий таблицу сопряженности из 4 чисел. Столь же многообразны попытки упорядочить возможные индексы и меры сопряженности. В недавней работе (Bossuyt, 2010) предлагается следующая классификация: 1) показатели точности (error-based) — чувствительность, специфичность и производные от них; 2) показатели информативности (information-based) — абсолютные риски при положительных и отрицательных результатах тестирования и отношения правдоподобий; 3) показатели сопряженности (association-based) — отношение шансов *OR*,

относительные риски и показатель «каппа» Коуэна. Эта классификация выглядит не слишком естественной. Фактически к показателям точности отнесены все оценки, связанные с исследованиями типа «случаи — контроли», а к показателям информативности отнесены оценки, характерные для когортных исследований. Абсолютные риски отнесены к показателям информативности, а их отношение (относительный риск) к показателям сопряженности. Кроме того, при $OR >> 1$ показатель «каппа» и относительные риски могут быть сколь угодно малы, а высокое отношение правдоподобий не гарантирует высокий уровень абсолютных рисков.

Мы будем придерживаться следующей нехитрой классификации: 1) тотальные показатели, т.е. зависящие от всей таблицы сопряженности; 2) условные показатели, при вычислении которых используются либо только строки, либо только столбцы таблицы сопряженности. Во втором случае мы будем всячески подчеркивать симметричность ситуации — каждому показателю по строчкам соответствует аналогичный показатель по столбцам. Фактически это единственный способ не запутаться в многообразии возможных характеристик теста. Кроме того, подобная классификация продиктована структурой данных в «генетике предрасположенностей». Генетики редко имеют возможность провести популяционное исследование генотипов и вычислить вероятность совмещения генетического маркера и заболевания. Обычно удается оценить лишь условные вероятности носительства маркера при наличии заболевания (исследования по схеме «случаи—контроль»), либо вероятности развития заболевания при условии носительства маркера (когортные исследования с целевым выбором носителей маркера). Эти два варианта соответствуют вычислениям условных показателей по столбцам либо по строкам.

Пусть совместное распределение вероятностей встречаемости маркера **M** и заболевания **D** задано в виде стандартной таблицы (матрицы) сопряженности $2 \times 2$

$$P = \begin{pmatrix} P(M,D) & P(M,\overline{D}) \\ P(\overline{M},D) & P(\overline{M},\overline{D}) \end{pmatrix} \quad (1)$$

с нормировкой $P(M,D) + P(M,\overline{D}) + P(\overline{M},D) + P(\overline{M},\overline{D}) = 1$. Здесь мы предполагаем, что бинарные случайные величины **M** и **D** принимают значения: $\mathbf{M} \in \{M, \overline{M}\}$ и $\mathbf{D} \in \{D, \overline{D}\}$. Под маркером $M$ понимается «предрасполагающий» генотип (аллель, гаплотип), сопряженный с заболеванием $D$. Под $\overline{M}$ имеется в виду совокупность альтернативных генетических вариантов. $\overline{D}$ означает отсутствие заболевания.

Введем обозначения для маргинальных сумм:

$p_D = P(M,D) + P(\overline{M},D)$ — распространенность заболевания, и

$p_M = P(M,D) + P(M,\overline{D})$ — популяционная частота маркера. Легко проверяемое тождество

$P(M,D) - p_M p_D = P(\overline{M},\overline{D}) - (1-p_M)(1-p_D) = p_M(1-p_D) - P(M,\overline{D})$

обусловливает возможность представления исходной матрицы $P$ в виде:

$$P = \begin{pmatrix} p_M p_D & p_M(1-p_D) \\ p_D(1-p_M) & (1-p_M)(1-p_D) \end{pmatrix} + \Delta \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad (2)$$

В этой сумме первая матрица соответствует случаю независимых случайных величин **M** и **D**, а вторая — добавка, возникающая за счет их взаимодействия.

Далее мы будем рассматривать исключительно случай $\Delta \geq 0$ (положительная связь между маркером и заболеванием). Случай $\Delta < 0$ соответствует замене $M \leftrightarrow \overline{M}$.

Величина $\Delta = P(M, D) - p_M p_D$ является одной из возможных (но редко используемых) мер сопряженности **M** и **D**. Определения наиболее распространенных показателей статистической сопряженности перечислены в таблице 1.

*Таблица 1*

**Показатели сопряженности маркера (M) и заболевания (D) для матрицы сопряженности *P***

| **Тотальные показатели сопряженности** | | | | |
|---|---|---|---|---|
| *OR* | $\dfrac{P(M,D)P(\overline{M},\overline{D})}{P(M,\overline{D})P(\overline{M},D)}$ | | | |
| *ACC* | $P(M,D) + P(\overline{M},\overline{D})$ | | | |
| $\Delta$ | $P(M,D) - p_M p_D$ | | | |
| *r* | $\Delta / \sqrt{p_D(1-p_D)p_M(1-p_M)}$ | | | |
| $\kappa_C$ | $2\Delta / (p_M(1-p_D) + p_D(1-p_M))$ | | | |
| **Условные показатели сопряженности** | | | | |
| | Истинные позитивы | Истинные негативы | Отношение рисков | Разность рисков |
| По столбцам | $SE = P(M \mid D)$ | $SP = P(\overline{M} \mid \overline{D})$ | $LR = SE/(1-SP)$ | $SE + SP - 1$ |
| По строчкам | $PPV = P(D \mid M)$ | $NPV = P(\overline{D} \mid \overline{M})$ | $RR = PPV/(1-NPV)$ | $PPV + NPV - 1$ |

*SE* (Sensitivity) — чувствительность, т.е. вероятность наличия маркера у субъекта с болезнью; *SP* (Specificity) — специфичность, т.е. вероятность отсутствия маркера у субъекта без болезни; *LR* (Likelihood Ratio) — отношение правдоподобий; *RR* (Risk Ratio) — отношение рисков; *PPV* (Positive Predictive Value) — предсказательная ценность наличия маркера, т.е. вероятность наличия болезни у субъекта с маркером; *NPV* (Negative Predictive Value) — предсказательная ценность отсутствия маркера, т.е. вероятность отсутствия болезни у субъекта без маркера.

*Тотальные показатели ассоциирования*

К их числу относятся отношение шансов (*OR*), точность (*ACC*), коэффициент корреляции (*r*) и показатель «каппа» Коуэна (Cohen's kappa — $\kappa_C$) (Cohen, 1960).

Очевидное преимущество показателя *OR* состоит в его универсальности в смысле применимости к любой схеме исследований (случаи — контроли, когортные исследования). Остальные тотальные показатели могут непосредственно оцениваться лишь в популяционных исследованиях (без целевого выбора «только больные» или только «носители маркера»).

Интуитивно привлекательным для понимания является показатель *ACC* (другое название — *FC*, Fraction Correct (Mitchell, 2009a, b)), который определяется как доля случаев правильного срабатывания теста (след матрицы *P*). Строго говоря, *ACC* не является показателем сопряженности, поскольку $ACC > 0$ при $OR = 1$. Более того, при $p_M \approx p_D << 1$ величина *ACC* слабо зависит от *OR* и близка к 1 даже в отсутствие статистической взаимосвязи.

В показателях *r* и $\kappa$ фактически используется разность $\Delta = P(M,D) - p_M p_D$. Тем не менее, коэффициент корреляции *r* в генетических исследованиях предрасположенности практически не фигурирует, т. к. непосредственно оценивается лишь в популяционных исследованиях. Величина *r* в первую очередь отражает линейность взаимосвязи **M** и **D**, т. е. близость матрицы *P* к диагональному виду. При этом в отличие от *OR* коэффициент корреляции может не регистрировать ситуации (быть близким к нулю), в которых носительство маркера является лишь необходимым (либо только достаточным) условием заболевания.

Показатель $\kappa_C$ часто используют для проверки согласия между двумя способами диагностики или между мнениями двух диагностов. При этом считается, что согласие хорошее при $0{,}6 \le \kappa_C \le 0{,}8$ и отличное при $0{,}8 \le \kappa_C \le 1{,}0$ (Landis, Koch, 1977). На практике показатель $\kappa_C$ близок к коэффициенту корреляции, но всегда $\kappa_C \le r$ с равенством при $p_M = p_D$. Точнее

$$r = \kappa_C \frac{1}{2}\left(\sqrt{\frac{p_D(1-p_M)}{p_M(1-p_D)}} + \sqrt{\frac{p_M(1-p_D)}{p_D(1-p_M)}}\right) \ge \kappa_C.$$

*Условные показатели ассоциирования*

Условные показатели (нижняя часть табл. 1) можно вычислять по столбцам либо по строчкам матрицы сопряженности в зависимости от схемы исследования. Обычно используют условные вероятности появления истинно позитивных и истинно негативных результатов тестирования. В исследованиях по схеме «случаи—контроли» по столбцам можно непосредственно оценить чувствительность (*SE*) и специфичность (*SP*) теста. При проведении когортных исследований непосредственной оценке поддаются двойственные показатели по строкам: предсказательная ценность для положительных (*PPV*) и отрицательных (*NPV*) результатов диагностического теста (positive/negative predictive value). Для каждой пары показателей можно определить относительные риски (*LR* и *RR*), которые всегда меньше *OR*. Двойственность определения условных показателей обусловливает выполнение тождеств:

$$\frac{SE}{PPV} = \frac{SE - p_M}{PPV - p_D} = \frac{OR - LR}{OR - RR} = \frac{p_M}{p_D}.$$

Ясно, что $OR > RR > LR$ при $p_M > p_D$, и $OR > LR > RR$ при $p_M < p_D$.

Показатель *LR* называют отношением правдоподобий и часто обозначают как $LR_+$, имея в виду выполнение тождества

$$LR_+ = \frac{SE}{1-SP} = \frac{P(M \mid D)}{P(M \mid \overline{D})} = \frac{P(D \mid M)/(1 - P(D \mid M))}{p_D/(1-p_D)},$$

которое позволяет интерпретировать $LR_+$ как отношение апостериорных шансов заболеть после получения информации о носительстве маркера к априорным шансам заболевания до получения такой информации. При этом вводится аналогичный показатель для отрицательных результатов тестирования:

$$LR_- = \frac{1-SE}{SP} = \frac{LR_+}{OR} = \frac{P(D|\overline{M})/(1 - P(D|\overline{M}))}{p_D/(1-p_D)}.$$

Мы будем рассматривать лишь $LR \equiv LR_+$ в виду его двойственности к условному показателю *RR* (относительный риск).

Среднюю эффективность теста часто характеризуют разностью абсолютных рисков:

$$P(M \mid D) - P(M \mid \overline{D}) = SE - (1 - SP)$$

(в исследованиях «случаи—контроли»),

$$P(D \mid M) - P(D \mid \overline{M}) = PPV - (1 - NPV)$$

(в когортных исследованиях).

Легко видеть, что показатели средней эффективности являются коэффициентами наклона соответствующих линий регрессии:

$$SE + SP - 1 = \frac{P(M,D) - p_M p_D}{p_D(1-p_D)} = r\frac{\sigma_M}{\sigma_D} = b_{M|D} \qquad (3)$$

$$PPV + NPV - 1 = \frac{P(M,D) - p_M p_D}{p_M(1-p_M)} = r\frac{\sigma_D}{\sigma_M} = b_{D|M},$$

где, $\sigma_M^2 = p_M(1-p_M)$, $\sigma_M^2 = p_M(1-p_M)$, $b_{M|D}$ и $b_{D|M}$ — наклоны линий регрессии **M** на **D** и **D** на **M**, соответственно. Имеются в виду регрессии, которые вычисляются после перекодировки: $\overline{M}, \overline{D} \to 0; M, D \to 1$.

Очевидно, что показатель $b_{D|M}$, являясь коэффициентом наклона регрессии **D** на **M** и разностью абсолютных рисков, характеризует среднюю диагностическую эффективность маркера, т. е. возможность предсказывать индивидуальную предрасположенность к заболеванию

по результатам тестирования на носительство маркера. В отношении показателя $b_{M|D}$ в следующем разделе будет показано, что в некотором смысле этот показатель характеризует возможности маркера решать классификационные задачи. Ничего другого и не следовало ожидать: показатель $b_{M|D,}$ являясь коэффициентом наклона регрессии **M** на **D**, характеризует способность теста отличать выборки больных от выборок здоровых.

Ясно, что коэффициент корреляции является средним геометрическим условных показателей эффективности теста:

$$r = \sqrt{(SE + SP - 1)(PPV + NPV - 1)}\,.$$

Отметим также, что величины $b_{M|D}$ и $b_{D|M}$ часто называют индексом Юдена (Youden, 1950) и суммарным предсказательным индексом (*PSI*, predictive summary index) (Linn, Grunau, 2006) соответственно.

## ВЕРОЯТНОСТНЫЕ ИНТЕРПРЕТАЦИИ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ ТЕСТА

Показатели *PPV* и *NPV* (а также производный от них показатель *RR*) имеют очевидную и практически важную интерпретацию — это условные вероятности развития заболевания при носительстве или отсутствии маркера. Прогностическая ценность двойственных показателей по столбцам (*SE* и *SP*) представляется менее очевидной. В этой связи рассмотрим несколько возможных вероятностных интерпретаций показателей, связанных с $b_{M|D} = SE + SP - 1$.

При анализе эффективности количественных маркеров успешно используется зависимость *SE* от 1 — *SP* (ROC-кривая). Площадь под этой кривой (*AUC* — Area Under Curve) равна вероятности того, что у случайно выбранного субъекта с болезнью значение мерного признака будет выше, чем у случайно выбранного субъекта без данной болезни (Fawcett, 2006). Для бинарного маркера ROC-кривая является кусочно-линейной (рис. 1), при этом соответствующая площадь равна

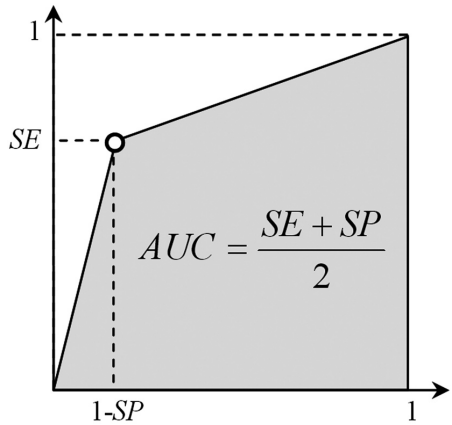



**Рис. 1.** «ROC-кривая» в случае бинарного теста. Площадь выделенной фигуры равна $AUC = (b_{M|D} + 1)/2 = (SE + SP)/2$

$$AUC = \frac{1}{2}(b_{M|D} + 1) = \frac{SE + SP}{2}\,.$$

Отметим также, что в работах, посвященных алгоритмам поиска межлокусных взаимодействий, показатель $(SE + SP)/2$ называют «балансовой точностью» (balanced accuracy, *BA*) (см., например, Winham et al., 2010).

В отношении показателя $(SE + SP)/2$ справедливо следующее

*Утверждение 1.* Пусть при тестировании одного больного и одного здорового индивидуума идентификация больного осуществляется следующим образом: больным объявляется носитель маркера, если результаты тестирования не совпадают, и больной выбирается случайно при совпадении результатов тестирования. Тогда вероятность правильной идентификации больного равна $AUC = (SE + SP)/2$.

*Доказательство.* Искомая вероятность равна

$$P_{2,1} = \frac{1}{2}P(M \mid D)P(M \mid \overline{D}) + \frac{1}{2}P(\overline{M} \mid D)P(\overline{M} \mid \overline{D}) + 1 \times P(M \mid D)P(\overline{M} \mid \overline{D}) + 0 \times P(\overline{M} \mid D)P(M \mid \overline{D}) = \frac{1}{2}SE \times (1 - SP) + \frac{1}{2}(1 - SE) \times SP + SE \times SP = \frac{1}{2}SE + \frac{1}{2}SP\,.$$

Утверждение доказано. Таким образом, величина $b_{M|D} = (AUC - 0{,}5)/0{,}5$ равна относительной добавке к вероятности 0,5 (случайное угадывание). В ROC-анализе принято считать, что маркер является хорошим классификатором при $AUC > 0{,}7$ и безусловно плохим при $AUC < 0{,}6$ (Swets, 1988).

Утверждение 1 допускает следующее обобщение, которое предлагает интерпретацию показателя $LR = SE/(1 - SP)$.

*Утверждение 2.* Пусть в группе из $n$ человек имеются один больной и $(n - 1)$ здоровых индивидуумов. Для обнаружения больного тестируются все члены группы, и если среди них обнаруживаются $k$ носителей маркера, то выбор больного среди них осуществляется случайным образом с вероятность $1/k$. Тогда вероятность правильной идентификации больного при тестировании группы равна

$$P_{n,1} = \frac{LR - SP^{n-1}(LR - 1)}{n} \underset{n \to \infty}{\sim} \frac{LR}{n}\,.$$

Иными словами, применение теста к группе из $n$ лиц увеличивает вероятность обнаружения больного в *LR* раз по сравнению со случайным угадыванием (которое возможно с вероятностью $1/n$).

*Доказательство.* Искомая вероятность равна

$$P_{n,1} = \frac{1}{n}(1-SE)SP^{n-1} +$$
$$+ SE\sum_{k=0}^{n-1}\frac{1}{k+1}C_{n-1}^{k}SP^{n-1-k}(1-SP)^{k} =$$
$$= \frac{1}{n}(1-SE)SP^{n-1} + SE\frac{1-SP^{n}}{n(1-SP)} =$$
$$= \frac{1}{n}\left(SP^{n-1} + \frac{SE}{1-SP}(1-SP^{n-1})\right) =$$
$$= \frac{1}{n}\left(LR - SP^{n-1}(LR-1)\right).$$

Утверждение доказано. При $n=2$ имеем формулу из Утверждения 1.

Широкое распространение получили показатели обратные к $b_{M|D}$ и $b_{D|M}$ в качестве оценок среднего числа тестов, проведенных до первого правильного срабатывания маркера. Этот подход заимствован из работ, оценивающих эффективность терапевтических методов, которые, как правило, являются когортными исследованиями (в качестве маркера $M$ выступает терапия $T$). В этих работах часто используется показатель $NNT = b_{D|T}^{-1} = (P(D \mid \overline{T}) - P(D \mid T))^{-1}$, который оценивает минимальную численность группы прошедших терапию, при которой число излеченных на одного больше, чем в такой же контрольной группе (Number Needed to Treat — число подлежащих воздействию). По аналогии для оценки эффективности использования маркеров различными авторами были предложены (см., Anonymous, 1996 и обсуждение: например, Mitchell, 2009 a, b):

- число подлежащих диагностированию (Number Needed to Diagnose) для исследований по схеме «случаи—контроль»

$$NND = b_{M|D}^{-1} = (P(M \mid D) - P(M \mid \overline{D}))^{-1} = \frac{1}{SE + SP - 1};$$

- число субъектов, необходимое для предсказания (Number Needed to Predict) для когортных исследований

$$NNP = b_{D|M}^{-1} = (P(D \mid M) - P(D \mid \overline{M}))^{-1} = \frac{1}{PPV + NPV - 1}.$$

При этом величину *NND* часто интерпретируют как среднюю численность выборки, которую необходимо протестировать, для обнаружения одного больного (Mitchell, 2009 a, b). Другие авторы полагают, что *NND* — это среднее число тестирований до момента любого правильного срабатывания теста (правильная идентификация больного или здорового) (Linn, Grunau, 2006). Нам представляется, что обе интерпретации ошибочны. Контрпример дает матрица $P$ вида

$$P = 10^{-6}\begin{pmatrix} 1 & 0 \\ 99 & 999900 \end{pmatrix}.$$

В этом случае $SE=0{,}01$ и $SP=1$ (все носители маркера больны). Тогда $NDD=100$, хотя для обнаружения больного при помощи маркера нужно в среднем провести $1/p_M = 10^6$ тестов. При этом доля случаев правильного срабатывания теста практически равна единице ($ACC = 0{,}999901$).

В отношении показателя $NND = b_{M|D}^{-1}$ можно утверждать лишь следующее. Пусть в единицу времени на носительство маркера проверяются один больной и один здоровый человек. Тогда среднее время ожидания события «число носителей маркера среди больных больше, чем среди здоровых» равно $b_{M|D}^{-1} = (SE + SP - 1)^{-1}$. При этом вероятность того, что в выборке больных число носителей маркера всегда больше, чем в выборке здоровых равна $SE + SP - 1$.

Аналогично, при рассмотрении растущей выборки носителей маркера среднее время ожидания события «число больных среди носителей маркера выше, чем среди свободных от маркера» равно

$$NNP = b_{D|M}^{-1} = (PPV + NPV - 1)^{-1}.$$

## ЗАВИСИМОСТЬ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ ОТ РАСПРОСТРАНЕННОСТИ ЗАБОЛЕВАНИЯ И ПОПУЛЯЦИОННОЙ ЧАСТОТЫ МАРКЕРА

Современные базы данных позволяют получать априорную информацию о частотах встречаемости возможных генов-маркеров ($p_M$) наряду с данными о распространенности изучаемого заболевания ($p_D$). В поисковых ассоциативных исследованиях прогностическая эффективность генетического теста будет существенно зависеть от популяционной частоты выбранного для исследования маркера. В этой связи необходимо четко представлять характер зависимости всех показателей эффективности теста от $p_M$ и $p_D$ при различных уровнях ассоциирования (значениях *OR*).

В этом разделе мы приведем сводку формул, описывающих зависимость показателей эффективности бинарного теста от трех независимых параметров: *OR*, $p_M$ и $p_D$. Начнем с вычисления ключевого показателя $\Delta = P(M,D) - p_M p_D$. Величина $\Delta$ вычисляется из определения *OR*:

$$OR = \frac{(p_M p_D + \Delta)((1-p_M)(1-p_D) + \Delta)}{(p_D(1-p_M) - \Delta)(p_M(1-p_D) - \Delta)},$$

откуда

$$\Delta = \frac{1}{2}\left( \frac{OR - \sqrt{\delta}}{OR - 1} - p_D p_M - (1 - p_D)(1 - p_M) \right) = \quad (4)$$

$$= \frac{4\sigma_M^2 \sigma_D^2 (OR - 1)}{(\sqrt{\delta} + 1)^2 - (p_M - p_D)^2 (OR - 1)^2}$$

где

$$\delta = 1 + 2(OR - 1)(p_D + p_M - 2p_D p_M) +$$

$$+ (OR - 1)^2 (p_D - p_M)^2.$$

Из вида выражения (4) ясно,что

$$ACC = (OR - \sqrt{\delta})(OR - 1)^{-1}.$$

Формуле для Δ можно также придать вид $\Delta = \Delta_{\max} \Delta$, где,

$$\Delta_{\max} = \min\{p_D(1 - p_M), p_M(1 - p_D)\} \text{ и}$$

$$\Delta' = 1 - \frac{2}{\sqrt{\delta} + 1 + |\, p_D - p_M \,|\,(OR - 1)}.$$

Напомним, что мы всюду рассматривает случай $\Delta \geq 0$, в котором всегда $OR \geq 1$. При $\Delta < 0$ выражение для $\Delta_{\max}$ имеет вид:

$$\Delta_{\max} = \min\{p_D p_M, (1 - p_M)(1 - p_D)\}.$$

В популяционной генетике величина Δ именуется «неравновесием по сцеплению», а Δ' «приведенным неравновесием по сцеплению» (Lewontin, Kojima, 1960; Slatkin, 2008). Отметим, что при $p_D < p_M$ величина Δ' совпадает с так называемым «популяционным атрибутивным риском» (*PAR*), который определяется (Levin, 1953) как

$$PAR = \frac{p_D - P(D \mid \overline{M})}{p_D} = \frac{NPV - (1 - p_D)}{p_D} =$$

$$= \frac{SE - p_M}{1 - p_M} = \frac{\Delta}{p_D(1 - p_M)} = \Delta'.$$

Выражения для *r* и Δ имеют достаточно громоздкий вид. В этой связи в таблице 2 мы приводим точные формулы для тотальных и условных показателей вместе с аппроксимациями для трех случаев: $OR \to 1$, $p_D \to 0$ и $p_M \to 0$. Таблица позволяет быстро оценивать прогностические возможности теста в крайних ситуациях. Например, при очень низкой распространенности заболевания ($p_D \to 0$) показатель *PPV* приблизительно равен $PPV \approx p_D OR(1 + p_M(OR - 1))^{-1} < p_D OR$. Это означает, что в случае редких заболеваний даже для очень «хорошего маркера» показатель *PPV* заведомо мал. Например, при $p_D = 10^{-4}$ и $OR = 100$ вероятность заболевания при носительстве маркера не превышает 1 %.

При $\sigma_D \sigma_M (\mathrm{OR} - 1) < 0{,}5$ хорошее приближение для Δ дает формула:

$$\Delta \approx \frac{\sigma_M^2 \sigma_D^2 (OR - 1)}{1 + (1 - ACC_0)(OR - 1)}, \quad (5)$$

где,

$\sigma_D^2 = p_D(1 - p_D)$, $\sigma_M^2 = p_M(1 - p_M)$,

$ACC_0 = p_D p_M + (1 - p_D)(1 - p_M)$ — доля «правильных тестирований» в отсутствие ассоциации. В этом приближении хорошо видна структура показателей эффективности маркера:

$$b_{M|D} \approx \frac{\sigma_M^2 (OR - 1)}{1 + (1 - ACC_0)(OR - 1)};$$

$$b_{D|M} \approx \frac{\sigma_D^2 (OR - 1)}{1 + (1 - ACC_0)(OR - 1)}.$$

В любом случае всегда справедливы неравенства $\Delta_{appr} \leq \Delta \leq 2\Delta_{appr}$, где $\Delta_{appr}$ — правая часть равенства (5).

Далее мы дадим качественное описание зависимостей показателей эффективности теста от распространенности заболевания и популяционной частоты маркера. Общий вид этих зависимостей показан на рисунках 2 и 3. Все условные показатели приведены за вычетом значений, характеризующих случай независимых **M** и **D**. Ясно, что во всех случаях зависимости для условных показателей представляются одной и той же поверхностью, которая от рисунка к рисунку зеркально отражается и поворачивается на 90°.

Качественная картина такова. Чувствительность теста слабо зависит от распространенности заболевания (монотонно убывает), но критично зависит от частоты встречаемости маркера (ярко выраженный максимум для редких заболеваний). Специфичность теста слабо зависит от распространенности заболевания (монотонно возрастает), но критично зависит от частоты встречаемости маркера для широко распространенных заболеваний (выраженный максимум). В отношении показателей *PPV* и *NPV* картина симметрично воспроизводится при замене местами $p_M \leftrightarrow p_D$.

Аналогичные зависимости для средних показателей эффективности $b_{M|D} = SE + SP - 1$ и $b_{D|M} = PPV + NPV - 1$ представлены на рисунке 3. Показатель $b_{M|D}$ слабо зависит от распространенности заболевания, но имеет максимум как функция частоты маркера. Напротив, показатель $b_{D|M}$ слабо зависит от частоты маркера, но имеет максимум как функция распространенности заболевания. На обоих рисунках гребень волны параллелен горизонтальной плоскости и расположен на высоте $(\sqrt{OR} - 1)(\sqrt{OR} + 1)^{-1}$. Отметим, что $\max b_{M|D} = \max b_{D|M}$ и совпадает с коэффициентом взаимосвязанности Юла (Yule's coefficient of colligation) (Yule, 1912). Следующее утверждение частично воспроизводилось многими авторами (см., например, King, Zeng, 2002).

*Таблица 2*

**Представление показателей эффективности маркера через $OR$, $p_M$ и $p_D$**

| | Точная формула | Приближенные формулы при | | |
|---|---|---|---|---|
| | | $OR \to 1$ | $p_D \to 0$ | $p_M \to 0$ |
| **Тотальные показатели сопряженности** | | | | |
| $ACC$ | $\frac{OR-\sqrt{\delta}}{OR-1}$ | $ACC_0$ | $1-p_M$ | $1-p_D$ |
| $\Delta$ | $\frac{1}{2}(ACC-ACC_0)$ | $\sigma_M^2\sigma_D^2(OR-1)$ | $\frac{\sigma_M^2 p_D(OR-1)}{1+p_M(OR-1)}$ | $\frac{\sigma_D^2 p_M(OR-1)}{1+p_D(OR-1)}$ |
| $r$ | $\frac{\Delta}{\sigma_D\sigma_M}$ | $\sigma_D\sigma_M(OR-1)$ | $\frac{\sigma_M\sqrt{p_D}(OR-1)}{1+p_M(OR-1)}$ | $\frac{\sigma_D\sqrt{p_M}(OR-1)}{1+p_D(OR-1)}$ |
| $\kappa_c$ | $\frac{2\Delta}{1-ACC_0}$ | $\frac{2\sigma_M^2\sigma_D^2(OR-1)}{1-ACC_0}$ | $\frac{2p_D(1-p_M)(OR-1)}{1+p_M(OR-1)}$ | $\frac{2p_M(1-p_D)(OR-1)}{1+p_D(OR-1)}$ |
| **Условные показатели сопряженности по столбцам** | | | | |
| $SE$ | $p_m+\frac{\Delta}{p_D}$ | $p_M+$ $+\sigma_M^2(1-p_D)(OR-1)$ | $\frac{p_M OR}{1+p_M(OR-1)}$ | $\frac{p_M OR}{1+p_D(OR-1)}$ |
| $SP$ | $1-p_m+\frac{\Delta}{1-p_D}$ | $1-p_M+$ $+\sigma_M^2 p_D(OR-1)$ | $1-p_M+$ $+\frac{\sigma_M^2 p_D(OR-1)}{1+p_M(OR-1)}$ | $1-\frac{p_M}{1+p_D(OR-1)}$ |
| $b_{M\|D}$ | $\frac{\Delta}{\sigma_D^2}$ | $\sigma_M^2(OR-1)$ | $\frac{\sigma_M^2(OR-1)}{1+p_M(OR-1)}$ | $\frac{p_M(OR-1)}{1+p_D(OR-1)}$ |
| $RR$ | $\frac{p_M+\frac{\Delta}{p_D}}{p_M-\frac{\Delta}{1-p_D}}$ | $OR-p_M(OR-1)$ | $\frac{OR}{1+p_M(OR-1)}$ | $OR-$ $-\frac{p_M OR(OR-1)}{1+p_D(OR-1)}$ |
| **Условные показатели сопряженности по строчкам** | | | | |
| $PPV$ | $p_D+\frac{\Delta}{p_M}$ | $p_D+$ $+\sigma_D^2(1-p_M)(OR-1)$ | $\frac{p_D OR}{1+p_M(OR-1)}$ | $\frac{p_D OR}{1+p_D(OR-1)}$ |
| $NPV$ | $1-p_D+\frac{\Delta}{1-p_M}$ | $1-p_D+\sigma_D^2 p_M(OR-1)$ | $1-\frac{p_D}{1+p_M(OR-1)}$ | $1-p_D+$ $+\frac{\sigma_D^2 p_M(OR-1)}{1+p_D(OR-1)}$ |
| $b_{D\|M}$ | $\frac{\Delta}{\sigma_M^2}$ | $\sigma_D^2(OR-1)$ | $\frac{\sigma_D^2(OR-1)}{1+p_D(OR-1)}$ | $\frac{p_D(OR-1)}{1+p_M(OR-1)}$ |
| $RR$ | $\frac{p_D+\frac{\Delta}{p_M}}{p_D-\frac{\Delta}{1-p_M}}$ | $OR-p_D(OR-1)$ | $OR-\frac{p_D OR(OR-1)}{1+p_M(OR-1)}$ | $\frac{OR}{1+p_D(OR-1)}$ |

Обозначения $\delta=1+2(OR-1)(1-ACC_0)+(OR-1)^2(p_D-p_M)^2$, $ACC_0=p_D p_M+(1-p_D)(1-p_M)$, $\sigma_M^2=p_M(1-p_M)$, $\sigma_D^2=p_D(1-p_D)$

*Утверждение 3.* При фиксированном *OR* максимально возможные значения средних показателей эффективности $b_{M|D}$ и $b_{D|M}$ равны

$$\max_{p_M, p_D}(SE + SP - 1) = \max_{p_M, p_D}(PPV + NPV - 1) = \frac{\sqrt{OR} - 1}{\sqrt{OR} + 1}$$

Для достижения максимума необходимо выполнение соотношения

$$(p_M - p_D)^2 OR = (1 - p_M - p_D)^2$$

*Доказательство.* Общеизвестно тождество

$$OR = SE \times SP(1 - SE)^{-1}(1 - SP)^{-1}.$$

Из соображений симметрии ясно, что максимум величины $b_{M|D} = SE + SP - 1 = SE \times SP(OR - 1)/OR$ достигается при $SE = SP$. Откуда $OR = SE^2/(1 - SE)^2$, а искомый максимум равен $2SE - 1 = (\sqrt{OR} - 1)/(\sqrt{OR} + 1)^{-1}$.

Аналогично максимум величины

$$b_{D|M} = PPV + NPV - 1 = PPV \times NPV(OR - 1)/OR$$

достигается при

$$PPV = NPV = \sqrt{OR}(\sqrt{OR} + 1)^{-1}$$

и равен

$$(\sqrt{OR} - 1)(\sqrt{OR} + 1)^{-1}.$$

Величина максимума не зависит от $p_M$ и $p_D$ (рис. 2). При этом max $b_{M|D}$ достигается при

$$p_M = (1 + p_D(\sqrt{OR} - 1))(\sqrt{OR} + 1)^{-1},$$

а max $b_{M|D}$ при

$$p_D = (1 + p_M(\sqrt{OR} - 1))(\sqrt{OR} + 1)^{-1}.$$

Утверждение доказано.

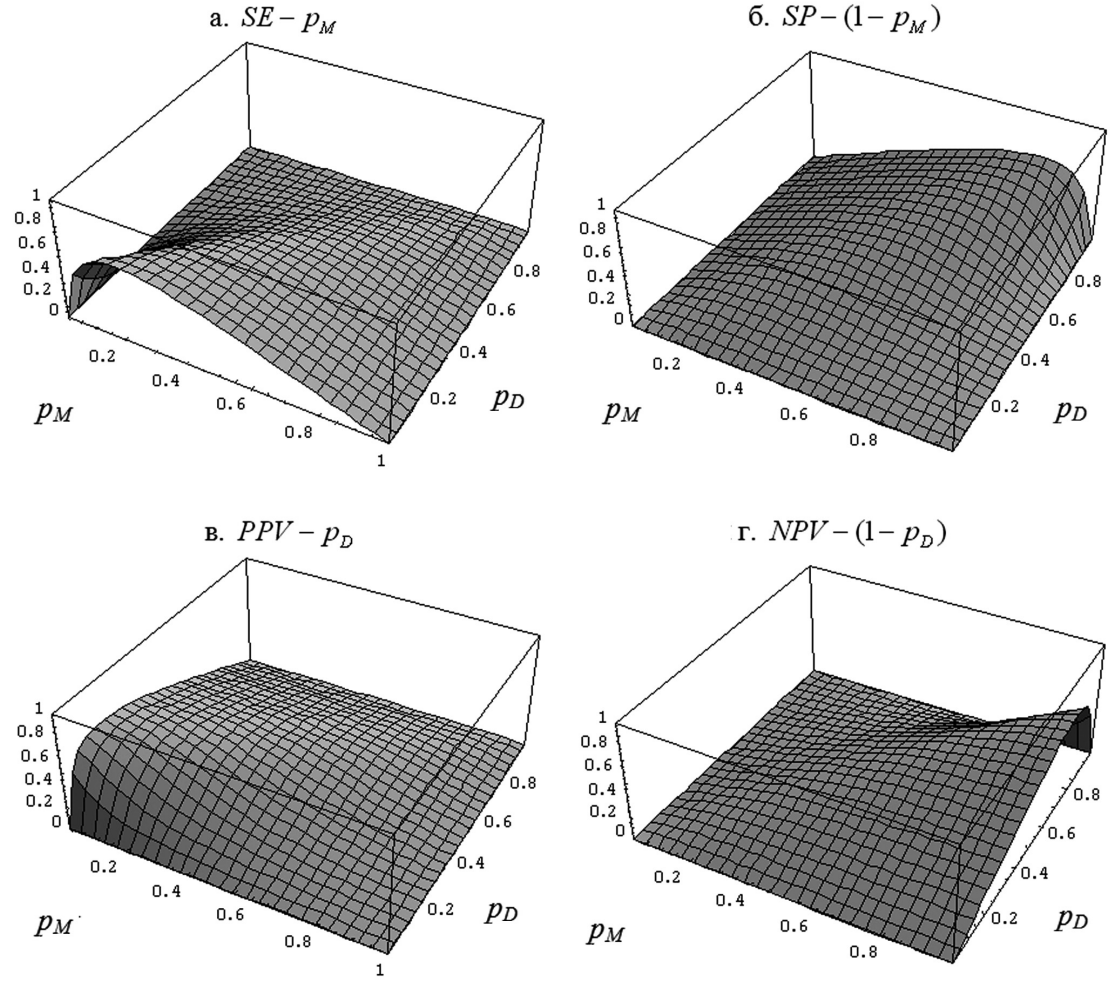



**Рис. 2.** Зависимость условных показателей эффективности от частоты встречаемости маркера ($p_M$) и распространенности заболевания ($p_D$) при $OR = 20$: а) показатель $SE - p_M$; б) показатель $SP - (1 - p_M)$ ; в) показатель $PPV - p_D$; г) показатель $NPV - (1 - p_D)$

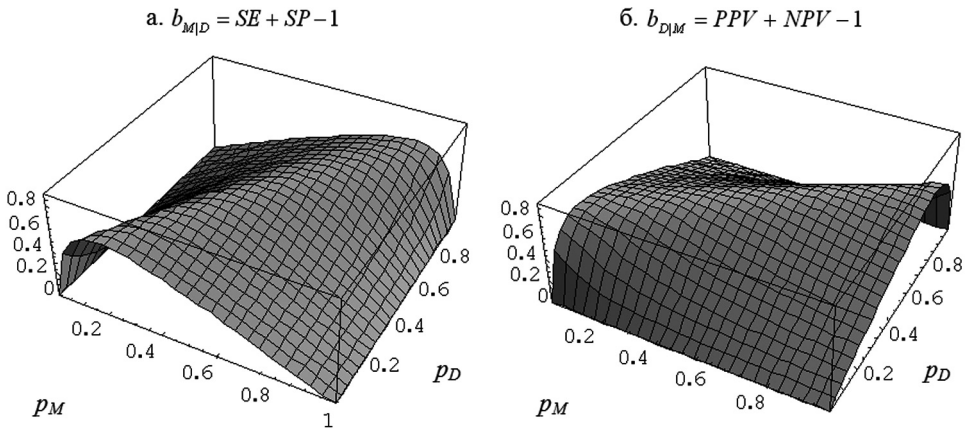



**Рис. 3.** Зависимость средних показателей эффективности от частот встречаемости маркера ($p_M$) и распространенности заболевания ($p_D$) при $OR = 20$: а) показатель $b_{M|D} = SE + SP - 1$; б) показатель $b_{D|M} = PPV + NPV - 1$. На обоих рисунках гребень волны параллелен горизонтальной плоскости и расположен на высоте $(\sqrt{OR} - 1)(\sqrt{OR} + 1)^{-1}$

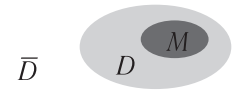
### *СЛУЧАЙ OR >> 1*

Предельное поведение показателей эффективности маркера из таблицы 1 при $OR \to \infty$ существенно зависит от соотношения между частотами встречаемости маркера и распространенности заболевания. Возможны две альтернативные ситуации, которые представлены в табл. 3.

Из таблицы, в частности, следует, что высокое значение *OR* и высокая статистическая значимость эффекта не всегда указывают на прогностическую эффективность маркера. Так при $OR \to \infty$ и $p_M >> p_D$ показатель диагностической эффективности $b_{D|M} = PPV = p_D / p_M < 1$, т.е. может быть сколь угодно мал по абсолютной величине. В противоположной ситуации ($p_M << p_D$) при $OR \to \infty$ за-

*Таблица 3*

**Предельное поведение показателей эффективности теста при $OR \to \infty$**

| | $p_M > p_D$ | $p_M < p_D$ |
|---|---|---|
| Вид таблицы сопряженности | $\begin{pmatrix} p_D & p_M - p_D \\ 0 & 1 - p_M \end{pmatrix}$ | $\begin{pmatrix} p_M & 0 \\ p_D - p_M & 1 - p_D \end{pmatrix}$ |
| Графическое представление |   |   |
| Причинность | M — необходимое условие заболевания | M — достаточное условие заболевания |
| *SE* | 1 | $p_M / p_D$ |
| *SP* | $(1 - p_M)/(1 - p_D)$ | 1 |
| *PPV* | $p_D / p_M$ | 1 |
| *NPV* | 1 | $(1 - p_D)/(1 - p_M)$ |
| *LR* | $(1 - p_D)/(p_M - p_D)$ | $\infty$ |
| *RR* | $\infty$ | $(1 - p_M)/(p_D - p_M)$ |
| $b_{M\|D}$ | $(1 - p_M)/(1 - p_D)$ | $p_M / p_D$ |
| $b_{D\|M}$ | $p_D / p_M$ | $(1 - p_D)/(1 - p_M)$ |
| *ACC* | $1 - (p_M - p_D)$ | $1 - (p_D - p_M)$ |
| $\Delta$ | $p_D(1 - p_M)$ | $p_M(1 - p_D)$ |
| *r* | $\sqrt{\frac{p_D(1 - p_M)}{p_M(1 - p_D)}}$ | $\sqrt{\frac{p_M(1 - p_D)}{p_D(1 - p_M)}}$ |

ведомо малы чувствительность теста и соответствующий показатель классификационной эффективности:

$$b_{M|D} = SE = p_M / p_D << 1.$$

## МАРКЕРЫ-КЛАССИФИКАТОРЫ И МАРКЕРЫ-ДИАГНОСТЫ

Для ситуаций, представленных в таблице 3 и 4, имеет смысл ввести специальные термины, отражающие специфику маркера. При высокой частоте встречаемости маркера ($p_M >> p_D$) и $OR >> 1$ имеем высокую чувствительность и высокую классификационную эффективность $b_{M|D}$, но низкую диагностическую значимость теста ($PPV \approx p_D / p_M$). Маркер рационально использовать для массового скрининга и профессионального отбора. По результатам тестирования можно отобрать группу заведомо здоровых людей (свободных от маркера). При этом для носителей маркера вероятность развития заболевания будет достаточно мала. По этой причине данную ситуацию можно обозначить как «маркер — классификатор». Типичным «маркером — классификатором» является маммография: $p_M = 0{,}04 >> p_D = 0{,}006$ при $OR = 200$, и $b_{M|D} = 0{,}84$. Тем не менее, вероятность наличия заболевания при положительных результатах тестирования достаточно низка — $PPV = 0{,}14$ и $b_{D|M} = 0{,}139$ (Banks et al. 2004).

Напротив, в случае редких маркеров ($p_M << p_D$) нет смысла проводить массовый скрининг — результаты будут заведомо «нулевые». Однако высокая диагностическая ценность теста $b_{D|M}$ при $OR >> 1$ позволяет его использовать в клинической практике при наличии дополнительных симптомов и показаний, например в случае неблагоприятной родословной. Подобный маркер можно назвать «маркером–диагностом». Пример «маркера–диагноста» демонстрируют данные по ассоциации полиморфизма *Leiden V* Arg506Gln с тромбозом вен (Folsom et al., 2002): $p_M = 0{,}07 << p_D = 0{,}32$ при $OR = 3{,}7$. Диагностическая ценность теста достаточно высока — $PPV = 0{,}61$ и $b_{D|M} = 0{,}31$. Однако как классификатор его использовать затруднительно: $b_{M|D} = 0{,}10$ из-за низкой чувствительности ($SE = 0{,}14$).

## ФОРМУЛЫ ДЛЯ ИССЛЕДОВАНИЙ «СЛУЧАИ–КОНТРОЛИ»

Как известно в исследованиях «случаи–контроли» невозможно напрямую оценить абсолютные (*PPV*, *NPV*) и относительные (*RR*) риски развития заболевания при наличии или отсутствии маркера. Однако с практической точки зрения часто именно эти оценки представляют наибольший интерес. Многие авторы предлагали для *RR* приблизительную формулу $RR \approx OR(1 - p_D + p_D OR)^{-1}$ (Zhang, Yu, 1998; Sistrom, Garvan, 2004). Легко видеть, что это есть точная нижняя оценка для *RR*. Точнее говоря, справедлива следующая цепочка неравенств

$$\frac{OR}{1 - p_D + p_D OR} \le RR \le OR - p_D(OR - 1) \le OR,$$

в которой левая и правая границы для *RR* соответствуют случаям $p_M = 0$ и $p_M = 1$ соответственно (рис. 4).

В принципе для полной реконструкции матрицы *P* необходимы три независимых показателя. Исследование по схеме «случаи–контроли» обеспечивает два из них: *SE* и *SP*. Третьим может быть распространенность заболевания ($p_D$), либо популяционная встречаемость маркера ($p_M$). В таблице 5 приведены оба типа оценок. Выбор между ними осуществляется в зависимости от того, какой из этих двух показателей мы считаем достоверно известным. В любом случае перед использованием формул из таблицы 4 необходимо проверить неравенство $SE > p_M > 1 - SP$, а также выполнение тождества:

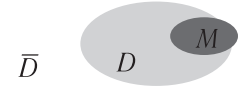
$$p_M = p_D SE + (1 - p_D)(1 - SP). \quad (6)$$

Интересно, что при наличии априорной информации о популяционной частоте встречаемости маркера для оценки относительного риска достаточно знать лишь частоту носителей маркера у больных:

$$RR = (1 - p_M) \times SE / (p_M \times (1 - SE)).$$

*Таблица 4*

**Качественное описание двух типов маркеров возможных при *OR >> 1***

| | **Маркер-классификатор** | **Маркер-диагност** |
|---|---|---|
| Графическое представление |   |   |
| Соотношение частот | $p_M >> p_D$ | $p_M << p_D$ |
| *SE*, $b_{M\|D}$ | Высокие | Низкие |
| *PPV*, $b_{D\|M}$ | Низкие | Высокие |
| Причинность | *M* — почти необходимое условие заболевания | *M* — почти достаточное условие заболевания |
| Использование | Массовый скрининг, профессиональный отбор | Наличие дополнительных симптомов, родословная |
| Примеры маркеров | Маммография | *BRCA, Leiden V* |

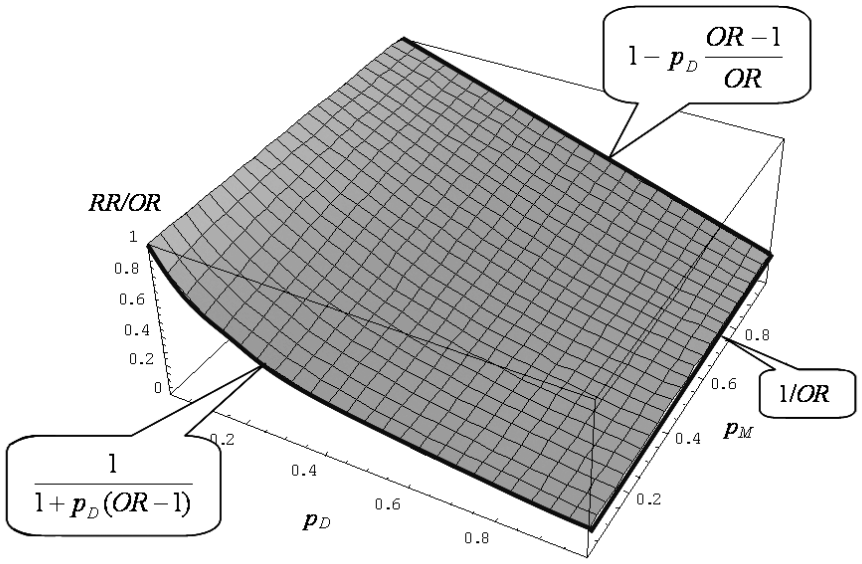



**Рис. 4.** Зависимость отношения $RR/OR$ от частоты встречаемости маркера ($p_M$) и распространенности заболевания ($p_D$) при $OR = 5$. Отношение $RR/OR$ слабо зависит от $p_M$, и монотонно убывает при увеличении $p_D$. При $p_M \to 0$ эта зависимость принимает вид $RR / OR = (1 - p_D + p_D OR)^{-1}$

*Таблица 5*

**Формулы для оценки показателей эффективности тестирования в исследованиях «случаи–контроли» через $SE$, $SP$ и $p_M$ (или $p_D$)**

| | Через $p_M$ | Через $p_D$ |
|---|---|---|
| *PPV* | $\frac{SE(SP-(1-p_M))}{p_M b_{M\|D}}$ | $\frac{p_D LR}{1+p_D(LR-1)}$ |
| *NPV* | $\frac{SP(SE-p_M)}{(1-p_M)b_{M\|D}}$ | $\frac{(1-p_D)SP}{SP-p_D b_{M\|D}}$ |
| $b_{D\|M}$ | $\frac{(SE-p_M)(SP-(1-p_M))}{\sigma_M^2 b_{M\|D}}$ | $\frac{\sigma_D^2 b_{M\|D}}{(SP-p_D b_{M\|D})(1-SP+p_D b_{M\|D})}$ |
| *RR* | $\frac{SE}{1-SE}\times\frac{1-p_M}{p_M}$ | $\frac{SE}{1-SE}\times\frac{SP-p_D b_{M\|D}}{1-SP+p_D b_{M\|D}}$ |
| $\Delta$ | $\frac{(SE-p_M)(SP-(1-p_M))}{b_{M\|D}}$ | $\sigma_D^2 b_{M\|D}$ |
| *r* | $\frac{\sqrt{(SE-p_M)(SP-(1-p_M))}}{\sigma_M}$ | $\frac{\sigma_D b_{M\|D}}{\sqrt{(SP-p_D b_{M\|D})(1-SP+p_D b_{M\|D})}}$ |
| *ACC* | $\frac{SE(SP-(1-p_M))+SP(SE-p_M)}{b_{M\|D}}$ | $p_D SE+(1-p_D)SP$ |
| *PAR* | $\frac{SE-p_M}{1-p_M}$ | $\frac{(1-p_D)b_{M\|D}}{SP-p_D b_{M\|D}}$ |
| Обозначения: $LR = SE/(1-SP)$, $b_{M\|D} = SE+SP-1$, $\sigma_D = \sqrt{p_D(1-p_D)}$, $\sigma_D = \sqrt{p_D(1-p_D)}$ | | |

## ЗАКЛЮЧЕНИЕ

Приведенные формулы позволяют оценить значения *OR* и частоты встречаемости маркера, гарантирующие высокую (или низкую) прогностическую эффективность соответствующего теста. В первую очередь следует выделить следующие три утверждения.

1. При $OR < 2{,}2$ маркер обладает заведомо низкой прогностической эффективностью во всех смыслах и при любых частотах встречаемости заболевания и маркера.
2. Маркер может быть хорошим классификатором, если $OR > 5{,}4$, при условии, что его популяционная частота достаточно высока ($p_M > 0{,}3$). На практике это означает, что указанным неравенствам должны удовлетворять нижние границы $100\,(1-\alpha)\,\%$-го доверительного интервала для оцениваемого значения $OR_L$, т. е. $OR_L < 2{,}2$ в первом случае и $OR_L > 5{,}4$ — во втором случае. Ранее близкие значения критических уровней наблюдаемых эффектов предлагались для относительных рисков ($RR < 2$ и $RR > 5$) (Ioannidis, 2006).
3. Даже при очень больших *OR* маркер является заведомо плохим классификатором ($AUC < 0{,}6$), если его популяционная частота низка ($p_M < 0{,}2\ p_D$). Аналогично, в силу неравенства $PPV < p_D\mathrm{OR}$ практически всякий маркер очень редкого заболевания обречен быть плохим диагностом.

   Действительно, из Утверждения 3 имеем

$$AUC = (b_{M|D} + 1)/2 < \sqrt{OR}\,/(\sqrt{OR} + 1).$$

Тогда, исходя из определения «плохого классификатора» ($AUC < 0{,}6$), получим $OR < 2{,}25$. В этом случае оба условных показателя средней эффективности ($b_{M|D}$ и $b_{D|M}$) и коэффициент корреляции ($r$) заведомо меньше

$$(\sqrt{2{,}25} - 1)/(\sqrt{2{,}25} + 1) = 0{,}2.$$

Далее, исходя из требования $AUC > 0{,}7$, получим $OR > 5{,}44$. При этом согласно Утверждению 3 максимум $b_{M|D}$ (а значит и *AUC*) достигается при

$$p_M = \frac{1 + p_D(\sqrt{OR} - 1)}{\sqrt{OR} + 1} > \frac{1}{\sqrt{OR} + 1} = \frac{1}{\sqrt{5{,}44} + 1} = 0{,}3.$$

Отметим также, что случай $AUC > 0{,}8$ возможен лишь при $OR > 16$ и $p_M > 0{,}2$.

Третье утверждение вытекает из формул, приведенных в таблице 3. При $OR >> 1$ и $p_M < p_D$ маркер является плохим классификатором, если

$$AUC = \frac{1}{2}\left(1 + \frac{p_M}{p_D}\right) < 0{,}6, \text{ или } p_M < 0{,}2\ p_D.$$

Итогом этого обсуждения является достаточно грустный вывод о низкой прогностической и классификационной эффективности результатов большинства опубликованных ассоциативных генетических исследований. Как правило, эти результаты укладываются в ситуацию из пункта 1, и не могут непосредственно использоваться в клинической практике. Тем не менее устойчиво воспроизводящиеся ассоциации даже при небольших *OR* могут указывать на участие определенных генов в становлении патологии, давая тем самым принципиально новую информацию о молекулярных механизмах заболевания.

Что же следует вычислять в случае редких удач — когда в исследовании по схеме «случаи-контроли» обнаруживается статистически высоко значимая ассоциация с высоким отношением шансов, например, $OR > 6$? Нам представляется, что, прежде всего, следует проверить полученные оценки *SE* и *SP* на согласованность с априорными данными по $p_M$ и $p_D$. Процедура проверки подразумевает два момента.

1) Проверка $p_M \in (1-SP,\ SE)$, т. е. принадлежности среднестатистических оценок популяционной частоты гена-маркера для данного этноса интервалу $(1-SP,\ SE)$, полученному в эксперименте.
2) Проверка оценки $p_D = (p_M - (1-SP))/b_{M|D}$, а именно ее соответствия общепринятым представлениям о распространенности данного заболевания.

Сильные отклонения от соотношения (6), подобно отклонениям от закона Харди—Вайнберга, могут указывать на ошибки генотипирования и/или идентификации фенотипа — заболевания. Возможны также эффекты, связанные с неоднородностью выборки. При удовлетворительном выполнении тождества (6) можно вычислить косвенные оценки *RR*, *PPV* и *NPV* согласно формулам из таблице 5. В результате будут получены оценки для обоих регрессионных коэффициентов, которые характеризуют прогностические возможности маркера. В целом, думается, что генетический маркер не безнадежен как классификатор, если $b_{M|D} > 0{,}4$, и как диагност, если $b_{D|M} > 0{,}4$. При этом редкий маркер может выступать только в качестве маркера—диагноста, и то лишь в случае широко распространенных заболеваний.

Все эти оценки будут иметь лишь предварительный характер. Очевидно, что всякую обнаруженную ассоциацию следует неоднократно верифицировать на независимых выборках. Кроме того, крайне желателен статистический анализ родословных, например в виде TDT-исследований (Spielman, 1994).

Работа выполнена при финансовой поддержке Минобрнауки России, ГК № 16.612.11.2061

## ЛИТЕРАТУРА

1. *Aly M., Wiklun F., Xu J. et al.,* 2011. Polygenic risk score improves prostate cancer risk prediction: results from the Stockholm-1 cohort study // European Urology. Vol. 60. P. 21—28.

2. Anonymous, 1996. How good is the test // Bandolier Journal. N 27. P. 2. http://www.medicine.ox.ac.uk/bandolier/painres/download/Bando027.pdf http://www.medicine.ox.ac.uk/bandolier/band27/b27–2.html.
3. *Banks E., Reeves G., Beral V.* et. al., 2004. Influence of personal characteristics of individual women on sensitivity and specificity of mammography in the Million Women Study: cohort study // BMJ. Vol. 329. N. 7464. P. 477–479.
4. *Bjartell A.,* 2011. Genetic markers and the risk of developing prostate cancer // European Urology. Vol. 60. P. 29–31.
5. *Bossuyt P.,* 2010. Clinical validity: Defining biomarker performance // Scandinavian Journal of Clinical & Laboratory Investigation. 70. P. 46–52
6. *Cohen J.,* 1960. A coefficient of agreement for nominal scales // Educational and Psychological Measurement. Vol. 20. P. 37–46.
7. *Fawcett T.,* 2006. An introduction to ROC analysis // Pattern Recognition Letters. Vol. 27. P. 861–874.
8. *Folsom A., Cushman M., Tsai M.* et al., 2002. A prospective study of venous thromboembolism in relation to factor V Leiden and related factors // Blood. Vol. 99. N. 9. P. 2720–2725.
9. *Ioannidis J.,* 2006. Commentary: Grading the credibility of molecular evidence for complex diseases // International Journal of Epidemiology. Vol. 35. P. 572–577.
10. *Jakobsdottir J., Gorin M.B., Conley Y.P.* et al., 2009. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers // PLoS Genet. Vol. 5. N 2. e1000337.
11. *King G., Zeng L.,* 2002. Estimating risk and rate levels, ratios, and differences in case-control studies // Statistics in Medicine. Vol. 21. P. 1409–1427.
12. *Kraft P., Wacholder S., Cornelis M.C.* et al., 2009. Beyond odds ratios — communicating disease risk based on genetic profiles // Nature Reviews Genetics. Vol. 10. P. 264–269.
13. *Kraemer H.C., Frank E., Kupfer D.J.,* 2011. How to assess the clinical impact of treatments on patients, rather than the statistical impact of treatments on measures // International Journal of Methods Psychiatric Research. Vol. 20. P. 63–72.
14. *Landis J.R., Koch G.G.,* 1977. The measurement of observer agreement for categorical data // Biometrics. Vol. 33. P. 159–174.
15. *Levin M.L.,* 1953. The occurrence of lung cancer in man // Acta Union International Contra Cancrum. Vol. 9. P. 531–541.
16. *Lewontin R.C., Kojima K.,* 1960. The evolutionary dynamics of complex polymorphisms // Evolution. Vol. 14. P. 458–472.
17. *Linn S., Grunau P.D.,* 2006. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests // Epidemiologic Perspectives & Innovations. Vol. 3: 11. http://www.epi-perspectives.com/content/3/1/11.
18. *Mitchell A.,* 2009a. How To: Implement a Screening Programme for Distress in Cancer Settings // Psycho-oncology.info. — Guide # 101. http://www.psycho-oncology.info/PG_implement_ajmitchell.pdf.
19. *Mitchell A.,* 2009b. How To: Analyse a Screening or Diagnostic Study // Psycho-oncology.info. — Guide # 104. http://www.psycho-oncology.info/PG_analyse_ajmitchell.pdf.
20. *Pepe M.S., Gu J.W., Morris D.E.,* 2010. The potential of genes and other markers to inform about risk // Cancer Epidemiology, Biomarkers & Prevention. Vol. 19. P. 655–665.
21. *Poste G.,* 2011. Bring on the biomarkers // Nature. Vol. 469. P. 156–157.
22. *Sistrom C.L., Garvan C.W.,* 2004. Proportions, odds, and risk // Radiology. Vol. 230. P. 12–19.
23. *Slatkin M.,* 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future // Nature Reviews Genetics. Vol. 9. P. 477–485.
24. *Spielman R.S., McGinnis R.E., Ewens W.J.,* 1994. Letter to the Editor: The transmission/disequilibrium test detects cosegregation and linkage // American Journal of Human Genetics Vol. 54. P. 559–560.
25. *Swets J.A.,* 1988. Measuring the accuracy of diagnostic systems // Science. Vol. 240. P. 1285–1293.
26. *Tan P.N., Kumar V., Srivastava J.,* 2004. Selecting the right objective measure for association analysis // Information Systems. Vol. 29. P. 293–313.
27. *Winham S.J., Slater A.J., Motsinger-Reif A.A.,* 2010. A comparison of internal validation techniques for multifactor dimensionality reduction // BMC Bioinformatics. Vol. 11:394. http://www.biomedcentral.com/1471–2105/11/394
28. *Youden W.J.,* 1950. Index for rating diagnostic tests // Cancer. Vol. 3. P. 32–35.
29. *Yule G.U.,* 1912. On the methods of measuring association between two attributes // Journal of the Royal Statistical Society. Vol. 75. P. 579–652.
30. *Zhang J., Yu K.F.,* 1998. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes // JAMA. Vol. 280. P. 1690–1691.

## THEORETICAL ANALYSIS OF THE PREDICTABILITY INDICES OF THE BINARY GENETIC TESTS

*Rubanovich A.V., Khromov-Borisov N.N.*

❀ **SUMMARY: A set of formulas for the indices of performance and predictive ability of the binary genetic tests is presented. Their dependence on disease prevalence and population frequency of a ge-**

**netic marker is characterized. It is shown that a marker with the odds ratio *OR* < 2.2 has an initially low prognostic efficiency in every sense and at any frequencies of the disease and the marker. A marker can be a good classifier, when *OR* > 5.4, but only when its population frequency is rather high (>0.3). The formulas are presented that allow to obtain indirect estimates of absolute and relative risk of the disease for the carrier of a marker in the case-control studies.**

❀ **KEY WORDS:** genetic association studies; odds ratio; area under curve (AUC); predictive genetic testing.

❀ Информация об авторах

**Рубанович Александр Владимирович** — зав. лаб. экологической генетики. ФГБУН «Институт общей генетики им. Н.И. Вавилова РАН». 119991, Москва, Губкина ул., д. 3.
E-mail: rubanovich@vigg.ru.

**Хромов-Борисов Никита Николаевич** — доцент. Кафедра физики, математики и информатики. Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова. 197022, Санкт-Петербург, ул. Льва Толстого, д. 6/8.
E-mail: Nikita.KhromovBorisov@gmail.com.

**Rubanovich Aleksandr Vladimirovich** — Head of Lab of ecological genetic in Vavilov Institute of General Genetics RAS. 119991, Moscow, Gubkin St., 3. Russia. E-mail: rubanovich@vigg.ru.

**Khromov-Borisov Nikita Nikolayevich** — associate professor. Department of Physics, Mathematics and Informatics in Saint-Petersburg State I. P. Pavlov Medical University. 197022, St.-Petersburg, Lev Tolstoy St., 6/8. Russia. E-mail: Nikita.KhromovBorisov@gmail.com.