УДК 616.07-036.8

Л.С. Файнзильберг

К вопросу о полезности диагностических методов в задачах скрининга

С позиций теории статистических решений развивается подход к оценке полезности диагностических методов (тестов) при массовых профилактических обследованиях населения. Доказаны утверждения, гарантирующие полезность диагностического теста с точки зрения уменьшения риска ошибочной диагностики. Получены оценки снизу граничных значений специфичности и чувствительности полезного теста.

An approach to estimation of usefulness of diagnostic methods (tests) under preventive inspections of the population is developed from the standpoint of the theory of statistical decisions. The statements guaranteeing the usefulness of a diagnostic test from the viewpoint of reducing the risk of erroneous diagnostics are proved. Lower-boundary values of specific character and sensitivity of the useful tests are obtained.

Розглянуто підхід до оцінки корисності діагностичних методів (тестів) для профілактичних обстежень населення, який грунтується на теорії статистичних рішень. Доведено твердження, які гарантують корисність діагностичного тесту з точки зору зменшення ризику невірної діагностики. Отримано оцінки знизу граничних значень специфічності та чутливості корисного тесту.

Введение. Диагностика имеет решающее значение во врачебной деятельности, поскольку только своевременно установленный диагноз позволяет выбрать адекватный метод лечения. Одним из основных направлений развития профилактической медицины является использование методов ранней диагностики заболеваний [1]. Для решения этой задачи необходимо проводить массовые обследования населения, в том числе людей, которые считают себя здоровыми и не обращаются к врачу. Подходы к реализации таких методов, получивших название скрининга (от англ. *screen* — отбирать, просеивать), изложены в ряде работ, в частности в [2–5].

Следует заметить, что идеальных методов диагностики не существует: часть больных признаются здоровыми (ложноотрицательный результат), а отдельные здоровые причисляются к больным (ложноположительный результат). Естественно, что при разработке того или иного метода стремятся, по возможности, минимизировать эти ошибки, или, что то же самое, повысить чувствительность и специфичность диагностического метода [6].

Понятно, что ложноотрицательный результат оставляет без дополнительного обследования и лечения больного человека, поэтому считается, что такие ошибки более опасны [7]. В то же время ложноположительные результаты часто наносят не меньший вред: здоровый человек может быть подвергнут небезопасному дополнительному обследованию (например, коронарографии при подозрении на органические поражения сосудов сердца, биопсии при подозрении на онкологическое заболевание и т.п.), не говоря уже о психологической травме, наносимой такими «диагнозами».

В связи с этим возникает вопрос: какие численные значения показателей чувствительности и специфичности диагностического метода (теста) следует считать приемлемыми для его рекомендации к практическому применению при скрининге того или иного заболевания?

В статье с позиций теории статистических решений проводится исследование данного вопроса и показывается, что не всякий тест, имеющий высокие показатели чувствительности и специфичности, может считаться полезным для задач скрининга.

Общая схема оценки полезности диагностического теста

Рассмотрим задачу скрининга группы, где есть и больные (класс V_1) и здоровые (класс V_2). При этом подразумевается, что V_2 включает не абсолютно здоровых людей, а лиц, у которых нет исследуемого заболевания.

Предположим, что для выявления больных используется диагностический тест, в соответствии с которым для конкретного пациента Z из обследуемой группы принимается решение в виде индикаторной функции

$$\delta = \begin{cases} 1, & \text{если принято решение } Z \in V_1; \\ 2, & \text{если принято решение } Z \in V_2. \end{cases} \quad (1)$$

Поскольку, как уже отмечалось, идеальных тестов не существует, решения (1) типа «Ты болен» или «Ты здоров» сопряжены с некоторым риском R ошибочной диагностики. Ясно и то, что априорные решения типа «Я болен» и «Я здоров», принимаемые без привлечения диагностического метода, также приводят к некоторому риску R_0 (априорному риску).

Поэтому для оценки полезности диагностического теста введем следующее определение.

Определение 1. Диагностический тест полезен для задач скрининга, если апостериорный риск меньше априорного, т.е. выполняется строгое неравенство

$$R < R_0. \quad (2)$$

Тогда оценку полезности диагностического теста следует проводить по схеме, представленной на рис. 1.

Задача скрининга с позиции теории статистических решений

Будем считать, что обследуемая группа репрезентативна в том смысле, что ее состав отражает генеральную совокупность, для которой известна априорная вероятность исследуемого заболевания $P(V_1)$. Соответственно $P(V_2) = 1 - P(V_1)$ будет обозначать априорную вероятность того, что пациент здоров.

В медицинской диагностике величина $P(V_1)$ обычно оценивается частотой заболевания (превалесом) [1]. Важной особенностью скрининга является низкий превалес больных в обследуемой группе. Так, например, согласно [1] превалес сахарного диабета составляет 2%, а рака молочной железы у женщин старше 50 лет — 1 % [8].

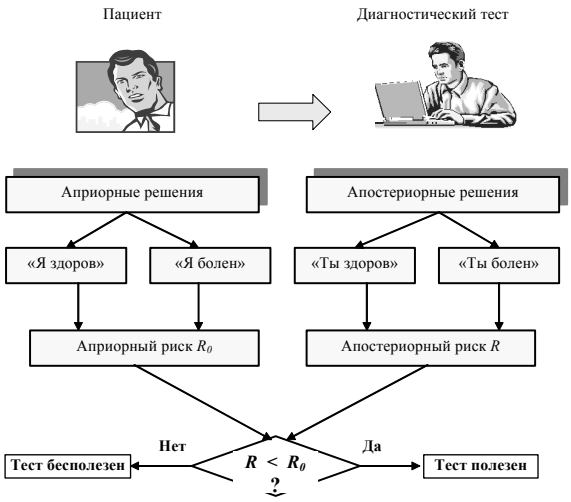

Рис. 1. Схема оценки полезности диагностического теста

Поставим задачу оценить условия, при которых выполняется строгое неравенство (2).

Пусть по некоторой репрезентативной группе испытуемых с известными диагнозами (экзаменационной группе) получены результаты тестирования, представленные матрицей решений (см. табл. 1). В этой таблице используются обозначения:

TP (*True Positive*) — число правильных диагнозов «Болен»;

TN (*True Negative*) — число правильных диагнозов «Здоров»;

FP (*False Positive*) — число здоровых, ошибочно признанных больными;

FN (*False Negative*) — число больных, ошибочно причисленных к здоровым.

Таблица 1. Результаты тестирования контрольной группы

Истинный диагноз	Результат тестирования	
	Решение — «Болен» (δ =1)	**Решение — «Здоров» (δ =2)**
Класс V_1 — «Болен»	***TP***	***FN***
Класс V_2 — «Здоров»	***FP***	***TN***

Согласно [6] «*Чувствительность*» теста определяется как доля (процент) больных, признанных больными по результатам тестирования, т.е.

$$\text{Чувствительность} = \frac{TP}{TP + FN}, \qquad (3)$$

а «*Специфичность*» теста, соответственно, — как доля здоровых, признанных после тестирования здоровыми, т.е.

$$\text{Специфичность} = \frac{TN}{TN + FP}. \qquad (4)$$

Очевидно, что если выборка репрезентативна, то чувствительность является оценкой величины $1 - P(e/V_1)$, где $P(e/V_1)$ — вероятность ошибочного отнесения больного к здоровым, а специфичность — оценкой величины $1 - P(e/V_2)$, где $P(e/V_2)$ — вероятность ошибочного отнесения здорового к больным[1].

Поскольку в общем случае эти ошибки не равнозначны, будем, как это принято в теории статистических решений [9, 10], характеризовать возможные потери платежной матрицей вида

$$\mathbf{L} = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix}, \qquad (5)$$

где L_{11} и L_{22}, L_{12} и L_{21} — потери, связанные, соответственно, с правильными решениями и ошибками первого и второго рода. Тогда средние потери при массовом обследовании лиц с неизвестными диагнозами будут определяться взвешенной суммой указанных потерь с учетом вероятностей их появления, т.е. средним риском

$$R = \sum_{k=1}^{2} \sum_{j=1}^{2} L_{kj} P(V_k, \delta = j), \qquad (6)$$

где $P(V_k, \delta = j)$ обозначает вероятность совместного выполнения двух случайных событий: пациент принадлежит k-му классу ($k = 1, 2$), а тест принял решение δ, отнеся его к j-му классу ($j = 1, 2$).

[1] В теории статистических решений такие ошибки называют ошибками первого и второго рода, или ошибками пропуска цели и ложной тревоги [9].

Для упрощения исследований допустим, что потери, связанные с правильными решениями, равны нулю, т.е. $L_{11} = L_{22} = 0$. Тогда если учесть, что по определению $P(V_k, \delta = j) = P(V_k)P(\delta = j/V_k)$, то при $j \neq k$ средний риск (6) можно представить в виде

$$R = L_{12} P(V_1) P(e/V_1) + L_{21} [1 - P(V_1)] P(e/V_2), \qquad (7)$$

или, в эквивалентной форме записи,

$$R = L_{12} \text{П}(1 - \text{Ч}) + L_{21}(1 - \text{П})(1 - \text{С}), \qquad (8)$$

где используются более привычные для задач медицинской диагностики термины: преваленс заболевания П, чувствительность Ч и специфичность С теста.

Таким образом, при известных П, Ч, С по формуле (8) можно количественно оценить средний риск применения диагностического теста, который в соответствии с (1) для каждого обследуемого пациента Z выдает одно из двух альтернативных решений: «Ты здоров» или «Ты болен».

Если же не проводить диагностического обследования, то каждый пациент может принимать одно из двух априорных решений: «Я здоров» (в этом случае риск будет равен L_{12}П) либо «Я болен» (риск равен $L_{21}(1 - \text{П})$). Легко видеть, что если выполняется условие

$$\text{П}(1 + \omega) < 1, \qquad (9)$$

где $\omega = L_{12} / L_{21}$ — соотношения потерь, связанных с ошибочным отнесением больного пациента к здоровым и наоборот, то с точки зрения минимума априорного риска для всех пациентов лучшим является решение «Я здоров», а при выполнении условия

$$\text{П}(1 + \omega) > 1 \qquad (10)$$

лучшим будет решение «Я болен».

На рис. 2 показана граница между областями указанных априорных решений. Видно, что при $\omega = 1$, когда потери от ошибок первого и второго рода считаются одинаковыми, априорное решение «Я здоров» следует принимать при П < 0,5. По мере же увеличения ω, когда возрастает «цена» ошибки отнесения больного к здоровым, такое решение следует

принимать при меньших значениях превалентса, что вполне логично. Например, при $\omega = 9$ априорное решение «Я здоров» следует принимать уже только при $П < 0{,}1$.

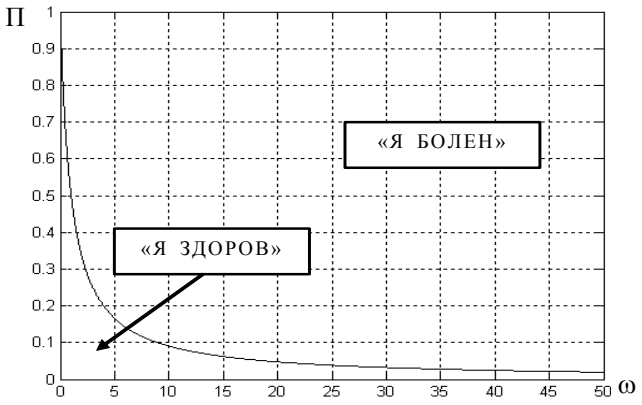

Рис. 2. Граница областей наилучших априорных решений

Таким образом, априорный риск R_0, фигурирующий в правой части неравенства (2), можно представить в виде

$$R_0 = \begin{cases} L_{12}П, \textit{ если } П(1+\omega) < 1 \\ L_{21}(1-П), \textit{ если } П(1+\omega) > 1. \end{cases} \qquad (11)$$

Подстановка (8) и (11) в (2) позволяет сформулировать следующее утверждение.

Утверждение 1. Диагностический тест полезен в смысле Определения 1 в том и только в том случае, когда при заданных превалентсе П, чувствительности Ч, специфичности С и соотношении ω потерь от ложноотрицательных и ложноположительных ошибок выполняется условие

$$(1-П)(1-С) < \omega ПЧ \qquad (12)$$

для области $П(1+\omega) < 1$

или условие

$$(1-П)С > \omega П(1-Ч) \qquad (13)$$

для области $П(1+\omega) > 1$.

Отсюда следует, что даже если диагностический тест обладает высокой специфичностью и чувствительностью, он оказывается бесполезным для проведения скрининга с точки зрения уменьшения риска ошибочной диагностики, если не выполняется условие (12) или условие (13).

Поскольку численные примеры часто бывают более убедительными, чем формальные рассуждения, рассмотрим пример.

Модельный пример 1. Предположим, что планируется провести скрининг заболевания, превалентс которого составляет 2%, т.е. $П = 0{,}02$. Для этого предлагается использовать тест с показателями $Ч = 0{,}9$ и $С = 0{,}9$. Пусть потери от ложноотрицательных и ложноположительных результатов составляют $L_{12} = 5$, $L_{21} = 1$.

В данном случае $П(1+\omega) < 1$ — значит, проверке подлежит условие полезности (12). Легко убедиться в том, что это условие не выполняется, т.е. тест абсолютно бесполезен для скрининга.

На рис. 3 представлены ожидаемые результаты тестирования группы из 10000 человек. При этом априорные потери, связанные с тем, что 200 больных будут не выявлены, составляет $5 \cdot 200 = 1000$ ед., в то время как суммарные потери, связанные с ложноположительными результатами диагностики для 980 здоровых пациентов и ложноотрицательными результатами для 20 человек, составят $1 \cdot 980 + 5 \cdot 20 = 1080$ ед. Следовательно, несмотря на высокие показатели чувствительности и специфичности, скрининг на основе такого теста только увеличит риск ошибочного диагностирования.

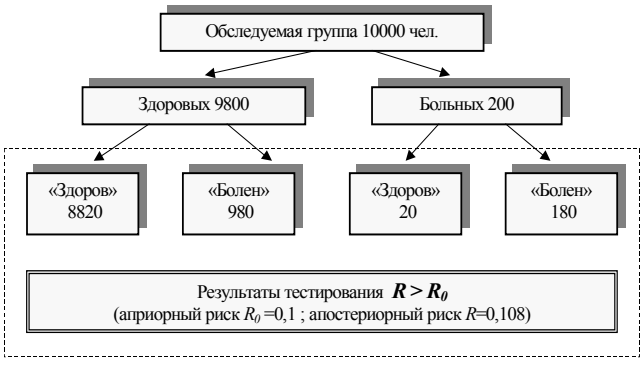

Рис. 3. Ожидаемые результаты скрининга при $П = 0{,}02$; $Ч = 0{,}9$; $С = 0{,}9$; $L_{12} = 5$; $L_{21} = 1$

Если изменить условия примера, положив $L_{12} = L_{21} = 1$, то увеличение риска в результате применения такого теста оказывается еще бо-

лее разительным: в этом случае ожидаемый средний риск скрининга составит $R = 0{,}1$, в то время как $R_0 = 0{,}02$.

Такой на первый взгляд парадоксальный результат имеет вполне конкретное объяснение. Дело в том, что если условие (12) не выполняется, то с точки зрения минимизации риска решение «Ты болен» является совершенно необоснованным.

Покажем это для случая, когда $L_{12} = L_{21} = 1$ и, как и раньше, $L_{11} = L_{22} = 0$. Как известно [11], при этом оптимальное решение соответствует максимуму апостериорных вероятностей распознаваемых классов.

Найдем апостериорные вероятности классов V_1 и V_2 при условии, что в результате тестирования принято решение «Ты болен». По формуле Байеса с учетом (1) их можно записать в виде

$$P(V_1/\delta=1)=\frac{P(V_1)P(\delta=1/V_1)}{P(\delta=1)}, \qquad (14)$$

$$P(V_2/\delta=1)=\frac{P(V_2)P(\delta=1/V_2)}{P(\delta=1)}, \qquad (15)$$

где $P(\delta = 1)$ — безусловная вероятность решения «Ты болен». На основании (14) и (15) заключаем, что такое решение будет обоснованным в том и только в том случае, когда

$$P(V_1)P(\delta=1/V_1) > P(V_2)P(\delta=1/V_2). \qquad (16)$$

С учетом принятых обозначений неравенство (16) можно представить так:

$$\text{П}\text{Ч} > (1-\text{П})(1-\text{С}). \qquad (17)$$

Однако если условие полезности (12) не выполняется (при заданных П, С, Ч и $\omega = 1$), то не выполняется и неравенство (17), т.е. $P(V_1/\delta=1) < P(V_2/\delta=1)$ и решение «Ты болен» будет неоптимальным.

В условиях данного примера апостериорные вероятности составляют $P(V_1/\delta=1)=0{,}155$ против $P(V_2/\delta=1)=0{,}845$ и $P(V_1/\delta=2)=0{,}003$ против $P(V_2/\delta=2)=0{,}997$. Следовательно, *для любого пациента* оптимальным является решение «Ты здоров», которое совпадает с априорным решением «Я здоров».

Заметим, что в соответствии с теоремами, доказанными в [12, 13], такое возможно только при $P(V_1) \neq P(V_2)$. Но именно этот случай как раз и характерен для задач скрининга.

Допустимые значения специфичности и чувствительности

В общем случае между специфичностью С и чувствительностью Ч диагностического теста не существует однозначной связи. Известно лишь то, что для некоторых алгоритмов диагностики с увеличением чувствительности снижается специфичность и наоборот. Такая ситуация характерна, например, для широко используемого в медицинской диагностике простейшего порогового решающего правила

$$\delta=\begin{cases}1, & \text{если } x > x_0;\\ 2, & \text{если } x < x_0,\end{cases}$$

где x— некоторый измеряемый физиологический параметр, а x_0 — пороговое значение. В выборе уровня скрининга (порога x_0) состоит основная проблема настройки такого рода диагностических тестов [1].

Покажем, как полученные условия (12) и (13) позволяют оценить значения С и Ч, гарантирующие полезность диагностического теста при заданных преваленсе П и соотношении потерь ω от ложноотрицательных и ложноположительных результатов. С этой целью, принимая во внимание, что $\text{П} \neq 1$, представим условие (12) в эквивалентной форме записи

$$C > \frac{1-\text{П}(1+\omega\text{Ч})}{1-\text{П}}. \qquad (18)$$

Поскольку всегда $\text{Ч} \leq 1$, то усиление (18) путем подстановки $\text{Ч} = 1$ приводит к оценке снизу граничного значения специфичности

$$\text{С}_{\text{гр}} = \frac{1-\text{П}(1+\omega)}{1-\text{П}} \qquad (19)$$

полезного теста для области $\text{П}(1+\omega) < 1$.

Из графика, приведенного на рис. 4, видно, что при малых значениях преваленса (П < 0,02),

характерных для задач скрининга ряда заболеваний, специфичность теста должна быть выше 90% уже при $\omega \le 5$, а при условии равенства потерь от ложноположительных и ложноотрицательных результатов — выше 98% .

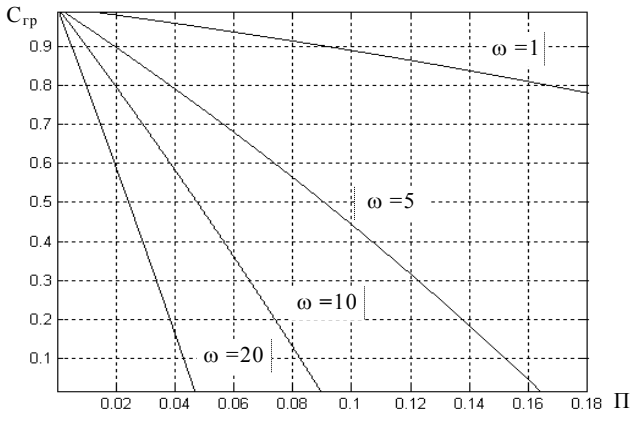

Рис. 4. Зависимость $C_{гр}$ от преваленса П при разных ω

Поскольку для реальных тестов Ч < 1, причем по мере снижения чувствительности растет число ложноотрицательных результатов, то, как следует из (19), для обеспечения полезности теста значение специфичности должно быть выше приведенных граничных величин, что еще более сужает область полезности теста.

В то же время, как видно из рис. 5, на котором представлены границы областей, удовлетворяющих условию (18), для низкого преваленса даже при ω = 1 снижение специфичности теста оказывается намного более критичным, чем уменьшение чувствительности. Так, например, при $П = 0{,}05$ и $\omega = 1$ допустимое значение С изменяется всего от 95,7% до 97,9% при существенном снижении чувствительности от 80% до 40% соответственно (рис. 6). Если же ω = 10, то при аналогичном изменении чувствительности и таком же преваленсе допустимое значение С находится в пределах 56,5–78,5 % (рис. 7).

Напомним, что соотношения (18) и (19) получены для случая $П(1+\omega) < 1$, который наиболее типичен для задач скрининга. В то же время известно [1, 7], что при диагностике особо опасных заболеваний соотношение потерь от ложноотрицательных и ложноположительных результатов иногда достигает $\omega \approx 400$. Для таких больших значений ω указанное условие может не выполняться даже при низком преваленсе (рис. 8).

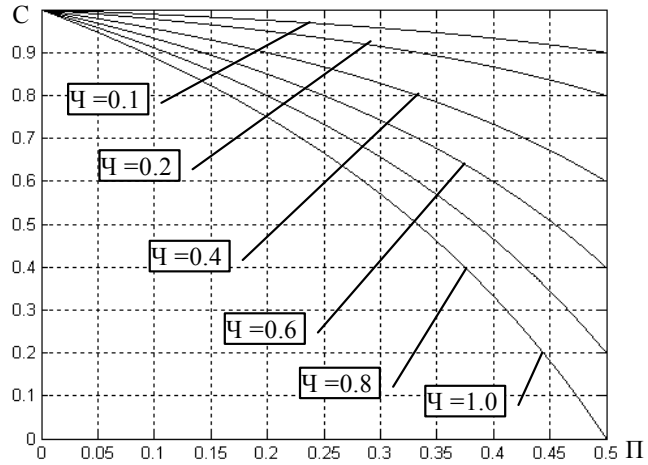

Рис. 5. Области полезности диагностического теста при ω = 1

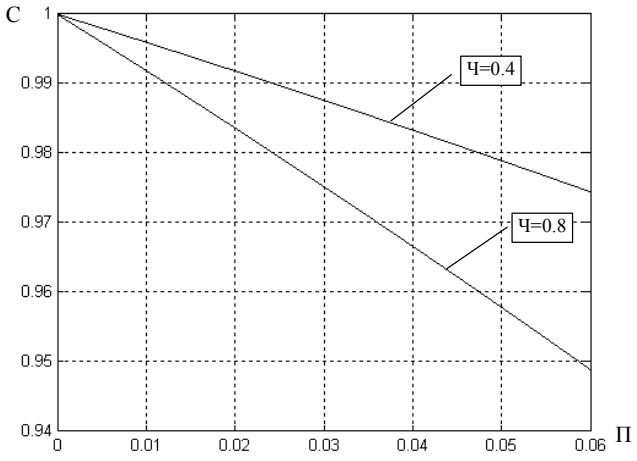

Рис. 6. Фрагмент областей полезности при ω = 1

Из условия (13) следует, что для области $П(1+\omega) > 1$ тест будет полезным только в том случае, когда

$$Ч > 1 - \frac{(1-П)С}{П\omega}. \qquad (20)$$

Усиление (20) путем подстановки С = 1 позволяет оценить снизу граничное значение чувствительности

$$Ч_{гр} = 1 - \frac{1-П}{П\omega} \qquad (21)$$

полезного диагностического теста для области $\Pi(1+\omega)>1$.

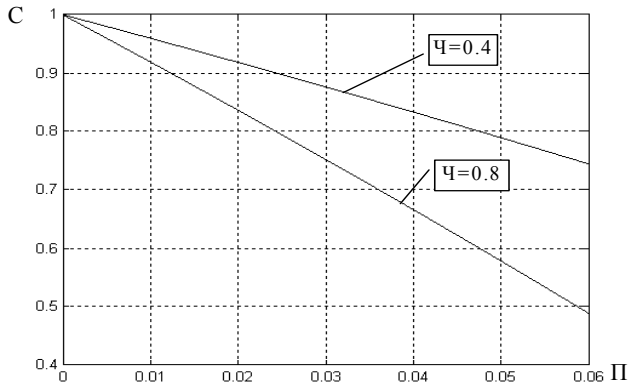

Рис. 7. Фрагмент областей полезности при ω = 10

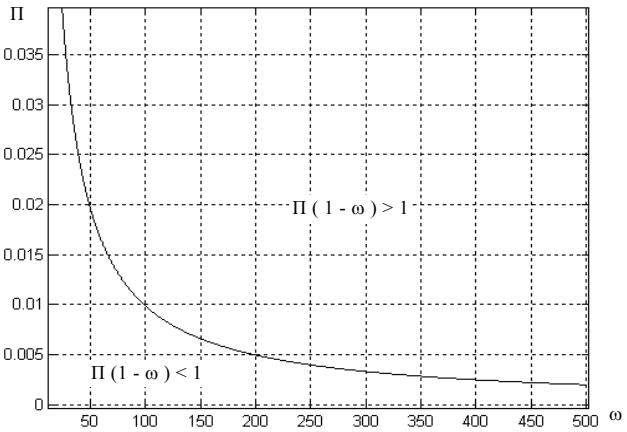

Рис. 8. Пояснения в тексте

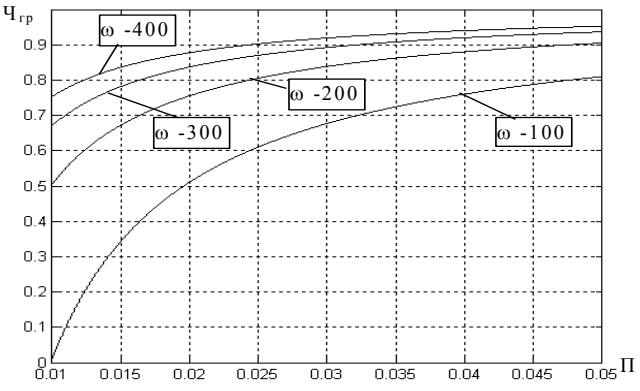

Рис. 9. Зависимость $Ч_{гр}$ от преваленса П при разных ω

На рис. 9 представлено семейство графиков для $Ч_{гр}$. Примечательно, что, в отличие от $С_{гр}$ (см. рис. 4), допустимая граница чувствительности уменьшается с уменьшением П и ω.

Для иллюстрации на рис. 10 показаны области полезности диагностического теста, построенные в соответствии с (20) при ω = 200. Заметим, что в данном случае снижение специфичности теста оказывается менее критичным, чем уменьшение чувствительности. Так, например, при П = 0,02 допустимое значение Ч может изменяться всего на 10% при снижении специфичности на 40%.

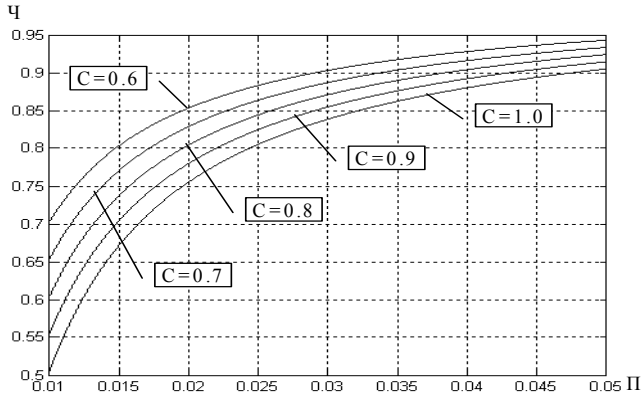

Рис. 10. Области полезности диагностического теста при ω = 200

Проведенные исследования позволяют в дополнение к Утверждению 1 сформулировать еще два не менее важных утверждения.

Утверждение 2. Если для области $\Pi(1+\omega)<1$ специфичность теста не превышает граничного значения, определяемого соотношением (19), т.е. $С \le \dfrac{1-\Pi(1+\omega)}{1-\Pi}$, то при *сколь угодно большой чувствительности* тест заведомо бесполезен.

Утверждение 3. Если для области $\Pi(1+\omega)>1$ чувствительность теста не превышает граничного значения, определяемого соотношением (21), т.е. $Ч \le 1-\dfrac{1-\Pi}{\Pi\omega}$, то при *сколь угодно большой специфичности* такой тест заведомо бесполезен.

Заключение. Исследования показывают, что высокие показатели чувствительности и специфичности сами по себе еще не гарантируют полезность диагностического теста с точки зрения уменьшения риска (суммарных

потерь) от ошибочной диагностики при массовых профилактических обследованиях населения.

Доказано, что диагностический метод (тест) будет полезен в том и только в том случае, когда выполняются условия (12) и (13) Утверждения 1. Получены оценки снизу граничных значений специфичности и чувствительности полезного теста, которые в соответствии с (19) и (21) зависят исключительно от преваленса П исследуемого заболевания и соотношения ω потерь от ложноотрицательных и ложноположительных ошибок.

Показано, что для области $П(1+\omega)<1$ критичной является величина специфичности, которая в соответствии с Утверждением 2 должна превысить граничное значение (19), а для области $П(1+\omega)>1$ критичной будет величина чувствительности, которая в соответствии с Утверждением 3 должна превысить граничное значение (21).

Полученные результаты могут быть использованы при разработке медицинских информационных технологий и позволяют обоснованно подходить к выбору показателей эффективности диагностических методов, рекомендуемых для практического применения в задачах скрининга.

1. *Власов В.В.* Эффективность диагностических исследований. — М.: Медицина, 1988. — 256 с.
2. *Казначеев В.П.*, *Баевский Р.М.*, *Берсенева А.П.* Донозологическая диагностика в практике массовых обследований населения. — Л.: Медицина, 1980. — 207 с.
3. *Массовые* медицинские обследования. — М.: Медицина, 1975. — 117 с.
4. *Thorner R.M.*, *Russel J.T.* Principles and procedures in the evaluation of screening for disease. — Washington: Gov. Print. Office, 1961. — 24 p.
5. *Blumberg M.S.* Evaluating health screening procedures // Operat. Res. — 1957. — **5**, N 3. — P. 351–360.
6. *Yerushalmy J.* Statistical problems in assessing methods of medical diagnosis with special reference to X-ray techniques // Publ. Health Rep. — 1947. — **62**, N 10. — P. 1432–1449.
7. *Ластед Л.* Введение в проблему принятия решений в медицине. — М.: Мир, 1971. — 282 с.
8. *Griner P.F.*, *Mayewski R.J.*, *Mushin A.I.*, *Greenland P.* Selection and interpretation of diagnostic tests and procedures // Ann. Intern. Med. — 1981. — **94**, N 4(2). — P. 553–600.
9. *Васильев В.И.* Распознающие системы: Справочник. — К.: Наук. думка, 1983. — 422 с.
10. *Горелик А.Л.*, *Скрипник В.А.* Методы распознавания. — М.: Высшая школа, 1977. — 220 с.
11. *Ковалевский В.А.* Методы оптимальных решений в распознавании изображений. — М.: Наука, 1976. — 328 с.
12. *Файнзильберг Л.С.* Оценка полезности признаков при решении задач диагностики в статистической постановке // Математические машины и системы. — 1998. — № 1. — С. 57–64.
13. *Fainzilberg L.S.* Why Relevant Features May Be Unuseful in Statistical Recognition of Two Classes? // Proc. of the 13th Intern. Conf. on Pattern Recognition (ICPR'96). — Viena (Austria), 1996. — P. 730–734.

Поступила 1.08.2002

Тел. для справок: 266-1154 (Киев)