

How to Read a Paper

The basics of evidence-based medicine

FIFTH EDITION

Trisha Greenhalgh

Department of Primary Care and Population Sciences
University College London Medical School
Highgate Hill
London

WILEY

BMJ|Books

Триша Гринхальх

ОСНОВЫ ДОКАЗАТЕЛЬНОЙ МЕДИЦИНЫ

Перевод с английского

**Под редакцией
акад. РАН И.Н. Денисова,
К.И. Сайткулова,
В.П. Леонова**

4-е издание, переработанное и дополненное



Москва
ИЗДАТЕЛЬСКАЯ ГРУППА
«ГЭОТАР-Медиа»
2018

Глава 8

Статьи, посвященные исследованиям диагностических и скрининговых тестов

8.1. Пример с десятью подсудимыми

Если вы незнакомы с методами оценки диагностических исследований, а математические объяснения (например, «давайте примем это значение за $x...$ ») у вас не вызывают энтузиазма, вам может помочь следующий пример. Десять подсудимых, обвиняемых в убийстве, ожидают решения судьи. Из них только трое действительно совершили преступление, семь человек невиновны. Судья рассматривает каждое дело и признает шесть человек виновными в преступлении. Из них двое являются истинными убийцами, а четверо осуждены неправильно. Один убийца остается на свободе.

Эту информацию можно представить в виде таблицы (табл. 8.1). Обратите внимание, что «истина» (т.е. совершал человек убийство или нет) отражена в верхнем горизонтальном ряду таблицы, а вердикт судьи (который может отражать, а может и не отражать истину) — в левом вертикальном столбце.

Таблица 8.1. Исходы процесса над 10 подозреваемыми в убийстве

		Истина	
		Убийца	Не убийца
Вердикт судьи	«Виновен»	Справедливо осуждены 2 человека	Несправедливо осуждены 4 человека
	«Невиновен»	Несправедливо оправдан 1 человек	Справедливо оправданы 3 человека

Вы должны видеть, что эти значения, если они являются для данного судьи типичными, отражают ряд его характеристик.

1. Судья правильно выявляет двух из каждых трех истинных убийц.

2. Он правильно оправдывает трех из каждых семи невинных людей.

3. Если этот судья счел человека виновным, вероятность того, что это действительно убийца, равна одной трети.

4. Если этот судья счел человека невиновным, вероятность того, что он действительно невиновен, равна трем четвертям.

5. В пяти случаях из десяти судья выносит правильное решение.

Эти пять характеристик составляют соответственно чувствительность, специфичность, прогностическую ценность положительного результата теста, прогностическую ценность отрицательного результата теста и индекс точности судьи. Остальная часть этой главы посвящена этим пяти характеристикам применительно к диагностическим (или скрининговым) тестам в сравнении с «истинным» диагнозом, или «золотым стандартом». В разделе 8.4 описывается шестая, более сложная (но очень полезная) характеристика диагностического исследования — отношение правдоподобия. (После прочтения главы обратитесь к данному ее разделу. К тому времени вы должны определить, что отношение правдоподобия положительного вердикта судьи в приведенном примере равно 1,17, а отрицательного вердикта — 0,78. Если не можете, не волнуйтесь — о том, что такое отношение правдоподобия, даже многие выдающиеся врачи не имеют понятия.)

8.2. Сравнение диагностического теста с «золотым стандартом»

Один рабочий рассказал мне, как он, почувствовав недавно жажду, обратился к своему врачу для исследования на диабет, которым страдают члены его семьи. Медсестра попросила его собрать образец мочи и опустила в него специальную полоску. Полоска окрасилась в зеленый цвет, что, по-видимому, означало, отсутствие в моче сахара (глюкозы). Медсестра сказала, что это свидетельствует об отсутствии диабета.

Мне было трудно объяснить рабочему, что результат исследования вовсе не свидетельствует об отсутствии диабета, как и судебное решение «виновен» не говорит о том, что человек *обязательно* является убийцей. Диабет, согласно определению ВОЗ, это уровень глюкозы в крови выше 7 ммоль/л натощак или выше 11,1 ммоль/л через 2 ч после теста на толерантность к глюкозе (очень неприятный тест, в ходе которого исследуемого просят выпить тошнотворно сладкий напиток со 100 г глюкозы и через 2 ч проводят исследование крови)¹. Эти по-

казатели должны быть зафиксированы два раза, если у человека нет симптомов, и только 1 раз, если у него есть типичные симптомы диабета (жажда, выделение большого количества мочи и т.д.).

Эти строгие критерии можно назвать *«золотым стандартом»* диагноза «диабет». Другими словами, если пациент отвечает критериям ВОЗ, он может считаться диабетиком, а при отсутствии критериев — нет (тем не менее эксперты ставят под сомнение такие категоричные заявления, как это; действительно, с момента первого издания этой книги значения в «золотом стандарте» диагностики диабета по уровню глюкозы крови изменились). Однако нельзя сказать то же самое в отношении опускания полосок в случайно взятый образец мочи. С одной стороны, у вас действительно может быть диабет, но почечный порог при этом высок; это означает, что почки сохраняют глюкозу гораздо лучше, чем у большинства людей, и уровень глюкозы в крови должен быть чрезвычайно высоким, чтобы она появилась в моче. Другой вариант — вы можете быть абсолютно здоровым человеком, но с *низким* почечным порогом; в этом случае глюкоза попадает в мочу, даже если в крови ее концентрация не повышена. Многие люди, страдающие диабетом, знают, что при этом заболевании часто исследование глюкозы в моче дает отрицательные результаты.

Тем не менее использование тест-полосок для скрининга на диабет имеет много преимуществ по сравнению с тестом на толерантность к глюкозе. Это дешевый, удобный, легкий в исполнении и интерпретации тест, дающий моментальный результат. В реальной жизни люди, как наш рабочий, могут отказаться от проведения неприятного теста на толерантность к глюкозе. Даже если он согласен, врач может решить, что симптомы не требуют проведения такого дорогостоящего и сложного исследования. Надеюсь, вы видите, что, хотя по исследованию мочи нельзя абсолютно точно поставить диагноз диабета, этот тест имеет определенные практические преимущества по сравнению с «золотым стандартом». По этой причине мы его и используем.

Чтобы объективно оценить, насколько информативно исследование глюкозы в моче для диагностики диабета, нам нужно отобрать группу людей (например, 100 человек) и провести у каждого из них оба теста: исследование мочи (скрининговый тест) и тест на толерантность к глюкозе («золотой стандарт»). Затем у каждого человека мы сможем увидеть, совпадают ли результаты скринингового теста с «золотым стандартом». Такое исследование известно как *оценка достоверности диагностического теста*. Мы можем выразить результаты этого исследования в виде таблицы (табл. 8.2) и подсчитать

различные характеристики теста (табл. 8.3) так, как мы это делали в отношении судьи в самом начале главы.

Если показатели различных характеристик теста (такие как чувствительность и специфичность) находятся в пределах разумного, мы можем говорить, что тест *достоверный* (см. ниже вопрос 7). Достоверность исследования глюкозы в моче для диагностики диабета изучалась Андерсон (Andersson) и соавт.², чьи данные я использовала в примере (табл. 8.4). На самом деле в оригинальное исследование были включены 3268 человек, из которых 67 либо отказались от исследования, либо по каким-то причинам это исследование не было им правильно проведено. Ради простоты я проигнорировала эти тонкости и выразила результаты в виде знаменателя (общее количество тестируемых) — 1000 человек.

Таблица 8.2. Таблица для выражения результатов оценки диагностического или скринингового теста (основано на Андерсон и соавт.)²

		Результат «золотого стандарта»	
		Заболевание есть a + c	Заболевания нет b + d
Результат скринингового теста	Положительный a + b	Истинно положительный a	Ложноположительный b
	c + d Отрицательный	c Ложноотрицательный	d Истинно отрицательный

На самом деле эти данные получены в ходе эпидемиологического исследования, направленного на определение распространенности диабета в популяции; оценка достоверности диагностического определения глюкозы в моче была побочным вопросом, параллельно решаемым в основном исследовании. Если бы оценка теста была основной целью исследования, группа включала бы гораздо больше больных диабетом (см. вопрос 2 в разделе 8.3)². Если вы обратитесь к оригинальной статье, то увидите, что «золотым стандартом» диагностики истинного диабета был не пероральный тест на толерантность к глюкозе, а гораздо менее стандартная серия наблюдений. Тем не менее этот пример соответствует своим задачам, поскольку он дает нам некоторые цифры, которые можно включить в уравнения, показанные в табл. 8.3. Важные характеристики исследования мочи на диабет можно считать следующим образом.

- Чувствительность = $a/a+c = 6/27 = 22,2\%$.
- Специфичность = $d/b+d = 966/973 = 99,3\%$.
- Прогностическая ценность положительного результата теста = $a/a+b = 6/13 = 46,2\%$.
- Прогностическая ценность отрицательного результата теста = $d/c+d = 966/973 = 97,8\%$.
- Индекс точности = $(a+d)/(a+b+c+d) = 972/1000 = 97,2\%$.
- Отношение правдоподобия положительного результата теста = $\text{чувствительность}/(1 - \text{специфичность}) = 22,2/0,7 = 32$.
- Отношение правдоподобия отрицательного результата теста = $(1 - \text{чувствительность})/\text{специфичность} = 77,8/99,3 = 0,78$.

Таблица 8.3. Характеристики диагностического теста, которые могут быть вычислены путем сравнения с «золотым стандартом» в исследовании по оценке достоверности теста

Характеристика теста	Другое название	Вопрос, на который отвечает данная характеристика теста	Формула (см. табл. 8.2)
Чувствительность (sensitivity)	Показатель истинной положительности (положителен при заболевании)	Насколько хорош тест для выявления людей, имеющих данное состояние?	$a/a + c$
Специфичность (specificity)	Показатель истинной отрицательности (отрицательный у здоровых)	Насколько хорош тест для правильного исключения людей, не имеющих данного состояния?	$d/b + d$
Прогностическая ценность положительного результата теста (positive predictive value)	Посттестовая вероятность наличия заболевания при положительном результате теста	Если у человека тест положительный, какова вероятность того, что у него действительно есть данное заболевание?	$a/a + b$
Прогностическая ценность отрицательного результата теста (negative predictive value)	Посттестовая вероятность отсутствия заболевания при отрицательном результате теста*	Если у человека тест отрицательный, какова вероятность того, что у него действительно нет данного заболевания?	$d/c + d$

Индекс точности (accuracy)	—	Какая часть всех тестов дала правильные результаты (т.е. истинно положительные и истинно отрицательные результаты по отношению ко всем)?	$(a+d)/(a+b+c+d)$
Отношение правдоподобия положительного результата (likelihood ratio of a positive test)	—	Насколько более вероятно то, что тест будет положительным у человека с заболеванием по сравнению со здоровым?	Чувствительность / (1 – специфичность)

* Посттестовая вероятность отрицательного результата теста равна (1 – прогностическая ценность отрицательного результата теста).

Таблица 8.4. Результаты оценки исследования глюкозы в моче для выявления диабета по сравнению с «золотым стандартом» — тестом на толерантность к глюкозе²

Результат теста на толерантность к глюкозе		
Результат исследования глюкозы в моче	Диабет есть 27 человек	Диабета нет 973 человека
Глюкоза обнаружена 13 человек	Истинно положительный 6	Ложноположительный 7
Глюкоза отсутствует 987 человек	Ложноотрицательный 21	Истинно отрицательный 966

Теперь вы видите, почему я не разделяла уверенности рабочего в том, что у него нет диабета. Чувствительность положительного теста на глюкозу в моче всего 22%; это означает, что тест не выявляет почти 4/5 истинных диабетиков. При наличии классических симптомов и семейного анамнеза изначальные шансы (претестовая вероятность) этого состояния у рабочего довольно высока и снижается только до 4/5 от этого значения (отношение правдоподобия отрицательного результата теста 0,78, см. раздел 8.4) после единичного отрицательного результата исследования глюкозы в моче. Учитывая симптомы, этому человеку необходимо пройти дополнительное исследование на диабет³. Обратите внимание, что, как показывают определения в табл. 8.4, если бы тест был положительным, тогда рабочему нужно было беспокоиться – хотя тест не

очень чувствительный (т.е. он не очень полезен для выявления людей с диабетом), он довольно специфичен (т.е. он полезен для исключения людей без заболевания)^{4, 5}.

Студенты часто путают различные параметры диагностического теста: чувствительность/специфичность и прогностическая ценность положительного/отрицательного результатов. Чувствительность и специфичность характеризуют *тест в целом*, а прогностическая ценность показывает, что *результаты данного теста означают для конкретного пациента*. Поэтому чувствительность и специфичность обычно больше используются эпидемиологами и специалистами в области общественного здоровья, чья ежедневная работа включает принятие решений о *популяциях*.

Скрининговая маммография (рентгенография молочной железы) может иметь чувствительность 80% и специфичность 90% для выявления рака молочной железы. Это означает, что тест выявляет 80% больных раком и исключает 90% женщин, не имеющих рака молочной железы.

Но представьте себя врачом или медсестрой, к которой приходит больная за результатами маммографии. Вопрос, который она вам задаст в случае положительного результата теста: «Какова вероятность, что у меня рак?», в случае отрицательного результата: «Какова вероятность того, что теперь я могу забыть о возможности этого заболевания?». Многие пациенты (и, к сожалению, многие врачи) полагают, что отрицательная прогностическая ценность теста равна 100%, т.е., если результат «нормальный», они думают, что вероятность заболевания равна нулю. Признания женщин, публикуемые в женских журналах («Мне сказали, что у меня рак, но затем тесты доказали, что врачи были неправы») — это примеры женщин, полагающих, что положительная прогностическая ценность теста равна 100%.

8.3. Десять вопросов, которые нужно задать о работе, оценивающей диагностический или скрининговый тест

При составлении этих вопросов я основывалась на трех основных источниках: «Руководство к чтению медицинской литературы»^{6, 7} и книга этих же авторов⁸, более свежая статья в журнале Американской медицинской ассоциации⁷ и рекомендации Дэвида Манты (David Mant) для оценки диагностических тестов⁹. Большинство проверочных листов в этой книге содержат не более чем общие практические советы для неопытных экспертов-новичков: если вы хотите ознако-

миться с исчерпывающим и точным набором критериев (объемом в 234 страницы) — проверочным листом QADAS [Quality in Diagnostic and Screening tests (Качество диагностических и скрининговых исследований)], рекомендую обратиться к недавно опубликованному обзору «Программа оценки технологий в здравоохранении» (Великобритания)⁸.

Вопрос 1. Подходит ли этот тест для моей практики?

Это то, что Лукас и его коллеги называют *полезностью* теста¹⁰. Даже если тест на 100% достоверный, точный и надежный, поможет ли он мне? Позволит ли он определить излечимое заболевание? Если да, предпочту ли я его тесту, который привык назначать? Могу ли я (мои пациенты, налогоплательщики, финансирующие медицинскую помощь) позволить это исследование? Согласятся ли мои пациенты его пройти? Изменит ли он вероятность возможного диагноза до такой степени, чтобы изменить план лечения? Если ответы на эти вопросы отрицательные, можно отклонить статью, не читая ее дальше аннотации или введения.

Вопрос 2. Сравнивался ли тест с «золотым стандартом»?

Сначала нужно задать вопрос, сравнивался ли диагностический тест в данном исследовании с чем-либо вообще. Некоторые авторы пишут статьи (в прошлом эти статьи даже публиковались), в которых не делалось ничего, кроме проведения нового теста на нескольких десятках пациентов. Такие упражнения могут давать различные результаты, но, конечно же, они не являются подтверждением того, что «высокие» результаты указывают на наличие выявляемого заболевания, а «низкие» результаты указывают на его отсутствие.

Далее следует проверить, заслуживает ли тест, использованный в исследовании, термина «золотой стандарт». Хороший способ оценки «золотого стандарта» — задать вопросы о полезности теста, перечисленные выше.

Для многих заболеваний абсолютного «золотого стандарта» диагностического теста, который определенно говорил бы о наличии или отсутствии заболевания, не существует. Неудивительно, что именно для этих состояний ученые наиболее активно ищут новые методы диагностики. В этом случае авторы часто разрабатывают комбинацию критериев, с которой и сравнивают новый оцениваемый тест. Один специфический момент, который нужно проверить, — не используется ли оцениваемый тест (или его разновидность) как способствующий определению «золотого стандарта».

Вопрос 3. Включало ли это исследование адекватный спектр пациентов?

Если вы оценивали новое исследование на холестерин у 100 здоровых студентов медицинской школы мужского пола, вы не сможете сказать, каковы будут результаты у женщин, детей, пожилых лиц, лиц с заболеваниями, связанными с повышенным уровнем холестерина, и даже у тех, кто никогда не бывал в медицинской школе. Конечно, исследователи не бывают настолько наивны, чтобы для оценки теста формировать такую *смещенную* группу. Однако в одной статье приводятся данные, что только в 27% опубликованных исследований четко определен состав пациентов с точки зрения их возраста, пола, симптомов и/или тяжести заболевания и специфических критериев отбора.

Определение состава участников исследования и спектра заболевания имеет большое значение, если результаты теста могут быть перенесены в другие условия. Диагностический тест может быть более чувствительным у женщин по сравнению с мужчинами или у молодых по сравнению с пожилыми. По этим же причинам, как отмечают Д. Сзекетт и соавт., группа, на которой проверяется тест, должна включать лиц с легкой и тяжелой формами заболевания, леченых и нелеченых, а также лиц с другими похожими состояниями.

В то время как чувствительность и специфичность теста постоянны вне зависимости от распространенности состояния, положительная или отрицательная прогностическая ценность во многом зависит от распространенности. Поэтому врачи общей практики скептически (часто правильно) относятся к полезности тестов, разработанных исключительно в условиях специализированной медицинской помощи, при которой тяжесть заболевания обычно выше (см. раздел 4.2). Поэтому же хороший *диагностический* тест (обычно используемый, когда у пациента имеются некоторые симптомы, предполагающие определенное заболевание) не обязательно является хорошим *скрининговым* тестом (обычно используется у людей без симптомов, исходящих из популяции с более низкой распространенностью заболевания).

Вопрос 4. Избегалась ли систематическая ошибка вследствие неполного проведения «золотого стандарта»?

Это легко проверить. Вопрос означает: «Каждый ли участник, получавший новый диагностический тест, также получал «золотой стандарт», и наоборот?». Потенциальная систематическая ошибка в

работах, в которых «золотой стандарт» проводился только у лиц с уже положительными результатами оцениваемого теста, очевидна. Кроме того, в такого рода смещениях есть и другие, более тонкие аспекты, выходящие за рамки этой книги. Подробное обсуждение этих вопросов можно найти в работе Райда (Reid) и соавт.¹¹

Вопрос 5. Избегалась ли систематическая ошибка вследствие ожидания?

Систематическая ошибка вследствие ожидания происходит, когда патолог или другой специалист, интерпретирующий диагностический материал, подсознательно находится под влиянием знания характеристик данного случая. В качестве примера можно привести знание врачом о наличии у пациента боли в груди при интерпретации электрокардиограммы. В контексте оценки диагностических исследований в сравнении с «золотым стандартом» вопрос означает: «Знали ли лица, интерпретировавшие один из тестов, каковы результаты другого теста у данного пациента?» Как я объясняла в разделе 4.5, все оценки должны быть «слепыми», т.е. человек, интерпретирующий результаты теста, не должен получать никаких намеков, какой результат может ожидать в данном конкретном случае.

Вопрос 6. Воспроизводим ли тест как у одного, так и между разными наблюдателями?

В определенном проценте случаев наблюдатель, проводящий одно и то же исследование два раза в разное время у пациента, характеристики которого не изменились, получает различные результаты. Это свойство в той или иной степени характерно для всех тестов. Однако тест с воспроизводимостью 99% находится в другой категории по сравнению с тестом, воспроизводимость которого равна 50%. Плохой воспроизводимости диагностического исследования может способствовать ряд факторов: техническое разрешение оборудования, вариабельность наблюдателей, проводящих тесты (например, при сравнении цветов с референтной цветовой шкалой), ошибки вычисления и т.д.

Обратитесь к разделу 4.5, чтобы вспомнить о проблеме согласия между различными наблюдателями. В интерпретации одного и того же результата два человека согласятся только в определенном проценте случаев, обычно выражаемом как показатель κ (каппа). Если же диагностическое исследование дает результаты в виде чисел (например, уровень холестерина в крови в ммоль/л), вопросов согласия между наблюдателями не возникает. Когда же тест включает интерпретацию рентгенограмм (пример с маммо-

графией в разделе 4.5) или вопросы о привычках, связанных с употреблением спиртных напитков¹⁰, важно подтвердить, что воспроизводимость между различными наблюдателями находится на приемлемом уровне.

Вопрос 7. Каковы характеристики теста, вытекающие из этого исследования?

Даже при соблюдении всех вышеуказанных стандартов диагностический тест может быть бесполезным, если он сам по себе недостоверен, т.е. его чувствительность, специфичность и другие ключевые характеристики слишком низкие. Именно к такой ситуации относится случай с определением глюкозы в моче для выявления диабета (см. раздел 8.2). В конце концов, если тест дает ложноотрицательные результаты примерно в 80%, он скорее вводит врача в заблуждение, нежели помогает подтвердить диагноз тогда, когда заболевание действительно имеется.

Для достоверности скринингового теста не существует абсолютных четких значений, поскольку приемлемые значения характеристик теста зависят от состояния, по поводу которого проводится скрининг. Мало кто будет «придираться» к тесту на цветовую слепоту с чувствительностью 95% и специфичностью 80%, но от этого заболевания никто не умирал. Скрининговый тест Гютри (Guthrie) на врожденный гипотиреозидизм, проводимый в Великобритании всем детям вскоре после рождения, имеет чувствительность 99%, но прогностическая ценность его положительного результата всего 6% (другими словами, тест выявляет практически всех детей с этим состоянием за счет высокого показателя ложноположительных результатов)¹¹. Выявить каждого ребенка с этим излечимым состоянием, который в противном случае может стать психически неполноценным, гораздо важнее, чем избавить сотни родителей от относительно небольшого стресса, связанного с повторным анализом крови у их ребенка.

Вопрос 8. Указаны ли доверительные интервалы для чувствительности, специфичности и других характеристик теста?

В разделе 5.5 объясняется, что доверительные интервалы, которые могут быть посчитаны практически для любого количественного результата, отражают возможный диапазон результатов, в пределах которого лежит истинное значение. Посмотрите снова на пример с судьей в разделе 8.1. Если бы он отнес еще одного убийцу к невиновным, чувствительность его решения упала бы с 67 до 33%, а прогностическая ценность положительного результата вердикта — с 33 до 20%. Эта огромная (и неприемлемая) зависимость от одного реше-

ния судьи, конечно, связана с тем, что мы оценивали решения судьи только по 10 случаям. Доверительные интервалы для характеристик этого судьи настолько широкие, что моя компьютерная программа отказывается их посчитать! Запомните: чем больше размер выборки, тем более узкий доверительный интервал. Поэтому на доверительные интервалы особенно важно обращать внимание, если размер выборки в статье небольшой. Формулу для расчета доверительных интервалов характеристик диагностического теста можно найти в прекрасной книге М. Гарднера и Д. Альтмана «*Статистика с уверенностью*»¹².

Вопрос 9. Выведен ли на основании результатов диапазон нормальных значений?

Если тест дает не дихотомические результаты, т.е. если он дает количественные значения, а не результат «да или нет», кто-то должен сказать, начиная с какого значения результат следует считать аномальным. Многие из нас проводили измерение собственного артериального давления. Мы хотим знать, нормальное ли у нас давление, а врач называет нам определенное значение, например 142/92. Если в качестве ориентира высокого артериального давления выбрано значение 140/90, мы окажемся в «аномальной» категории, хотя риск возникновения у нас проблем практически не отличается от такового у человека с давлением 138/88. Многие врачи в таких случаях вполне разумно говорят своим пациентам следующее: «Давление у вас не вполне нормально, однако оно не находится в опасной зоне. Через 3 мес нам нужно снова его проверить». Так или иначе, в некоторый момент врач должен принять решение: *это* давление требует лечения таблетками, а *другое* – нет¹³.

Определение зон относительного и абсолютного риска для непрерывных физиологических или патологических параметров – сложная наука, в которой следует принимать во внимание вероятность побочных исходов, на предотвращение которых направлено лечение. Более объективным этот процесс делает использование отношений правдоподобия (см. раздел 8.4). Увлекательное обсуждение различных возможных значений слова «нормальный» в диагностических исследованиях можно найти в книге Д. Скетта и соавт.¹⁴, на с. 59.

Вопрос 10. Интерпретируется ли данный тест в контексте других тестов в диагностическом поиске при данном состоянии?

В целом высокое артериальное давление мы лечим на основании только измерения давления (хотя мы стараемся полагаться на серию, а не на однократное измерение). Сравните это с последовательностью,

которую мы используем для диагностики стеноза коронарных артерий. Сначала мы отбираем больных с типичной историей стенокардии напряжения (боли в сердце при физических упражнениях). Затем мы проводим ЭКГ в покое, ЭКГ с физической нагрузкой и, в некоторых случаях, радионуклидное исследование для определения областей в сердце с недостаточным поступлением кислорода. Коронарную ангиографию (наиболее точное исследование стеноза коронарных артерий) большинству пациентов проводят только *после* того, как были получены аномальные результаты предварительных тестов.

Если вы возьмете 100 случайных людей и отправите их на коронарную ангиографию, этот тест может показать различные значения положительной и отрицательной прогностической ценности (и даже различную чувствительность и специфичность) по сравнению с популяцией больных лиц, в которой этот тест первоначально оценивался. Это означает, что различные аспекты достоверности коронарной ангиографии как диагностического исследования практически бессмысленны, пока эти цифры не выражаются в терминах их вклада в общий диагностический поиск.

8.4. Отношение правдоподобия

В вопросе 9 поднимается проблема определения диапазона нормы для непрерывной переменной. В этих обстоятельствах результаты теста предпочтительнее выражать не как «нормальные» или «патологические», а в терминах вероятности у больного заболевания, если результат теста достигает определенного уровня. Примером может служить использование простатоспецифического антигена (ПСА) для скрининга на рак предстательной железы. У большинства мужчин в крови определяется некоторое значение ПСА (например, 0,5 нг/мл); у большинства мужчин, страдающих раком предстательной железы, уровень ПСА очень высок (выше 20 нг/мл). Уровень ПСА 7,4 нг/мл может быть обнаружен как у абсолютно здорового мужчины, так и у человека с ранней стадией рака. Четкая граница между нормой и патологией отсутствует¹⁵.

Результаты оценки теста ПСА в сравнении с «золотым стандартом» (биопсия предстательной железы) можно использовать для составления серии таблиц типа табл. 8.2. В разных таблицах будут использоваться различные определения аномального результата ПСА для разделения всех лиц на «норму» и «патологию». Из этих таблиц мы можем получить различные отношения правдоподобия, связанные со значениями ПСА, находящимися в различных диапазонах. В этом случае, даже если значение ПСА у нашего пациента окажется в

«серой» зоне, мы сможем сказать: «Это исследование не подтвердило рак предстательной железы у больного, но оно повысило (или снизило) вероятность этого диагноза на x ». (Как я говорила в разделе 6.3, тест на ПСА не позволяет четко провести различие между наличием и отсутствием рака независимо от того, какое значение используется для отделения нормы от патологии. Другими словами, нет такого значения ПСА, которое давало бы высокое отношение правдоподобия для выявления рака.)¹⁶

Хотя отношение правдоподобия — наиболее сложный для вычисления аспект диагностического теста, он имеет огромную практическую ценность и становится все более предпочтительным способом выражения и сравнения пользы различных тестов. В своем руководстве Д. Сэккетт и соавт.⁶ объясняют, как отношение правдоподобия можно использовать для подтверждения или исключения конкретного диагноза. Например, если у пациента нет никаких симптомов, я знаю, что вероятность железодефицитной анемии у него составляет 5%, поскольку в популяции в среднем этим заболеванием страдает 1 человек из 20 [на языке диагностических исследований это означает, что претестовая (априорная) вероятность анемии у пациента равна распространенности этого состояния и составляет 0,05]¹⁷.

Рис. 8.1 представляет собой номограмму, адаптированную Д. Сэккеттом и соавт. на основе оригинального исследования Т. Фэгена (Т. Fagan)¹⁸, для определения посттестовой вероятности на основе претестовой вероятности (распространенности) и отношения правдоподобия. Линии А, В и С начинаются с претестовой вероятности 25% (распространенность курения среди взрослых лиц в Великобритании) и проходят через отношения правдоподобия соответственно 15, 100 и 0,015 — три различных теста для выявления курения¹⁹. Тест С фактически позволяет определить, что человек *не курит*, поскольку положительный результат этого теста ведет к посттестовой вероятности всего 0,5%.

И последнее. Как я говорила в начале главы, вы можете использовать диагностические тесты, не обращаясь к отношениям правдоподобия. Я сама долгие годы обходилась без них. Но если вы посвятите полдня освоению этого аспекта клинической эпидемиологии, ваше время будет потрачено не зря.

8.5. Правила клинического прогнозирования

В предыдущем разделе приводился довольно трудно воспринимаемый пример ПСА-теста, и было сделано заключение, что не су-

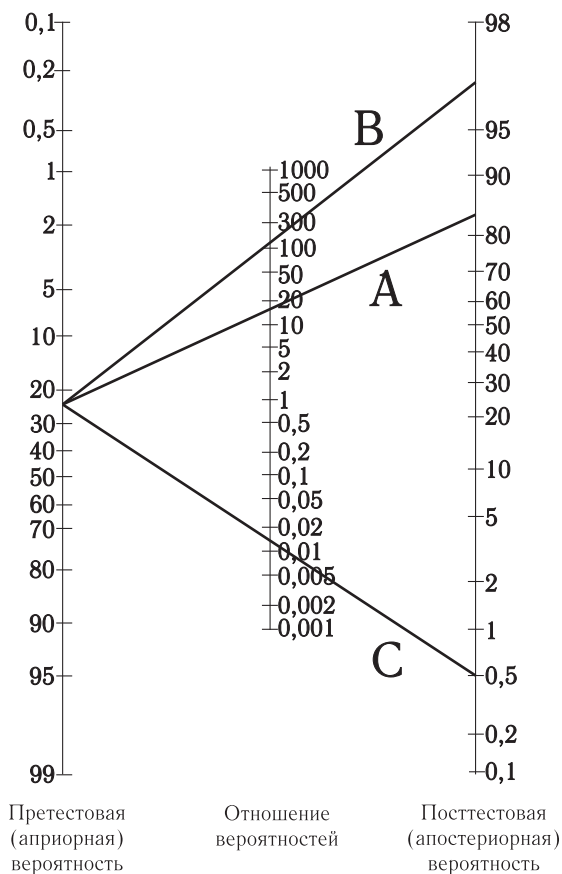


Рис. 8.1. Использование отношений правдоподобия для вычисления посттестовой вероятности курения

существует однозначного, четкого значения, которое надежно отличало бы «норму» от «патологии». Поэтому рекомендованный подход к оценке риска рака простаты у мужчин является комбинацией нескольких тестов, включая общую клиническую оценку и цифровое ректальное обследование¹⁶.

В общем и целом вы, наверное, можете понять, почему зачастую клиницисты стремятся использовать комбинацию нескольких диагностических тестов (включая клиническое обследование, анализ крови, рентгенографию и т.п.) для создания картины того, что случилось с пациентом. В то время как любой из тестов имеет размытые

границы между нормой и патологией, в комбинации они способны сделать диагностику более четкой. Так, например, женщине, которая обратилась с опухолью молочной железы, скорее всего предложат три разных теста, ни один из которых не является однозначным сам по себе: аспирационная диагностическая пункция, рентгенография (маммограмма) и ультразвук²⁰. В последнее время ученые начали дискуссию относительно того, повышает ли расшифровка компьютерной маммографии точность этой тройной комбинации²¹.

Общепринятая практика – делать несколько тестов и сравнивать их – является традиционным общим правилом в клинической практике, недавно доработанным в более структурированную форму Фальком и Фаеем (Falk and Fahey)²². Обследуя большие когорты пациентов с конкретными симптомами и тщательно протоколируя результаты всех клинических обследований и диагностических тестов, мы можем прийти к численным оценкам шансов того, что человек страдает (или близок к тому, чтобы заболеть) заболеванием X при наличии симптома A, физического признака B, диагностического теста C, и так далее – или любой их комбинации. Интерес к правилам клинического прогнозирования и их исследованию в последние годы рос высокими темпами, частично из-за того, что развитие информационных технологий означает, что очень большое количество пациентов может быть включено в базу данных клиницистами в разных центрах.

Как отмечают Фальк и Фаей (Falk and Fahey), существует три этапа развития правила клинического прогнозирования. Во-первых, правило должно быть разработано посредством установления независимого и комбинированного воздействия объяснимых переменных, таких как симптомы, признаки или диагностические тесты на заболевания. Во-вторых, эти объяснимые переменные должны быть оценены в разных популяциях. И в-третьих, должен быть анализ влияний – в идеале рандомизированное исследование, которое измеряет влияние применяемого правила в клинических условиях в отношении результата для пациента, клинического поведения, использования ресурса и так далее.

В качестве примера того, как правила клинического прогнозирования могут помочь нам проработать самые запутанные диагностические проблемы в здравоохранении, ознакомьтесь с этими работами о том, как делать прогноз относительно целесообразности направления на компьютерную томографию²³ ребенка с травмой головы; развивается ли ревматоидный артрит у пациента с ранним артритом²⁴, находится ли пациент, принимающий антикоагулянты, в группе достаточно низкого риска инсульта, чтобы прекратить их прием²⁵, и ка-

кая комбинация тестов лучше всего определит, действительно ли тяжело заболевший ребенок страдает чем-то серьезным²⁶.

Литература

1. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. Geneva: World Health Organization, 2006:1–50.

2. Andersson D, Lundblad E, Svärdsudd K. A model for early diagnosis of type 2 diabetes mellitus in primary health care. *Diabetic Medicine* 1993;10(2):167–73.

3. Friderichsen B, Maunsbach M. Glycosuric tests should not be employed in population screenings for NIDDM. *Journal of Public Health* 1997;19(1):55–60.

4. Bennett C, Guo M, Dharmage S. HbA1c as a screening tool for detection of type 2 diabetes: a systematic review. *Diabetic Medicine* 2007;24(4):333–43.

5. Lu ZX, Walker KZ, O'Dea K, et al. A1C for screening and diagnosis of type 2 diabetes in routine clinical practice. *Diabetes Care* 2010;33(4):817–9.

6. Jaeschke R, Guyatt G, Sackett DL, et al. Users' guides to the medical literature: III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA: The Journal of the American Medical Association – US Edition* 1994;271(5):389–91.

7. Guyatt G, Bass E, Brill-Edwards P, et al. Users' guides to the medical literature: III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA: The Journal of American Medical Association* 1994;271(9):703–7.

8. Guyatt G, Sackett D, Haynes B. Evaluating diagnostic tests. *Clinical epidemiology: how to do clinical practice research* 2006;424:273–322.

9. Mant D. Testing a test: three critical steps. *Oxford General Practice Series* 1995;28:183.

10. Lucas NP, Macaskill P, Irwig L, et al. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology* 2010;63(8):854–61.

11. Lu Y, Dendukuri N, Schiller I, et al. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. *Statistics in Medicine* 2010;29(24):2532–43.

12. Altman DG, Machin D, Bryant TN, et al. *Statistics with confidence: confidence intervals and statistical guidelines*. Bristol: BMJ Books, 2000.

13. Appel LJ, Miller ER, Charleston J. Improving the measurement of blood pressure: is it time for regulated standards? *Annals of Internal Medicine*. 2011;154(12):838–9.

14. Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston: Little, Brown and Company, 1985.

15. Holmström B, Johansson M, Bergh A, et al. Prostate specific antigen for early detection of prostate cancer: longitudinal study. *BMJ: British Medical Journal* 2009;339:b3537.

16. Barry M, Denberg T, Owens D, et al. Screening for prostate cancer: a guidance statement from the Clinical Guidelines Committee of the American College of Physicians. *Annals of Internal Medicine* 2013;158:761–9.

17. Guyatt GH, Patterson C, Ali M, et al. Diagnosis of iron-deficiency anemia in the elderly. *The American Journal of Medicine* 1990;88(3):205–9.

18. Fagan TJ. Letter: nomogram for Bayes theorem. *The New England Journal of Medicine* 1975;293(5):257.

19. Moore A, McQuay H, Muir Gray J. How good is that test – using the result. *Bandolier*. Oxford 1996;3(6):6–8.

20. Houssami N, Irwig L. Likelihood ratios for clinical examination, mammography, ultrasound and fine needle biopsy in women with breast problems. *The Breast* 1998;7(2):85–9.

21. Giger ML. Update on the potential of computer-aided diagnosis for breast cancer. *Future Oncology* 2010;6(1):1–4.

22. Falk G, Fahey T. Clinical prediction rules. *BMJ: British Medical Journal* 2009;339:b2899.

23. Maguire JL, Boutis K, Uleryk EM, et al. Should a head-injured child receive a head CT scan? A systematic review of clinical prediction rules. *Pediatrics* 2009;124(1):e145–54.

24. Kuriya B, Cheng CK, Chen HM, et al. Validation of a prediction rule for development of rheumatoid arthritis in patients with early undifferentiated arthritis. *Annals of the Rheumatic Diseases* 2009;68(9):1482–5.

25. Rodger MA, Kahn SR, Wells PS, et al. Identifying unprovoked thromboembolism patients at low risk for recurrence who can discontinue anticoagulant therapy. *Canadian Medical Association Journal* 2008;179(5):417–26.

26. Verbakel JY, Van den Bruel A, Thompson M, et al. How well do clinical prediction rules perform in identifying serious infections in acutely ill children across an international network of ambulatory care datasets? *BMC Medicine* 2013;11(1):10.