CAPÍTULO 4

Avaliação da Validade e da Confiabilidade dos Testes Diagnósticos e de Rastreamento

Para entender como uma doença é transmitida, como ela se desenvolve e para oferecer uma assistência em saúde apropriada e efetiva, é necessário distinguir as pessoas na população que têm a doença daquelas que não a têm. Isso é um importante desafio tanto na área clínica, onde o tratamento do paciente é o objetivo, como na área de saúde pública, onde os programas secundários de prevenção que envolvem a detecção precoce e a intervenção de doenças estão sendo considerados e onde os estudos etiológicos estão sendo conduzidos para estabelecer uma base para a prevenção primária. Portanto, a qualidade do rastreamento e dos testes diagnósticos é um tema fundamental. Independentemente se o teste for o exame físico, o raios X, o eletrocardiograma ou exames de sangue ou urina, ocorre a mesma questão: o quanto o teste é bom na separação de populações de pessoas com e sem a doença em questão? Este capítulo refere-se à questão de como nós avaliamos a qualidade do rastreamento e dos testes diagnósticos recentemente disponíveis para tomar decisões razoáveis com relação à sua utilização e interpretação.

VARIAÇÃO BIOLÓGICA DE POPULAÇÕES HUMANAS

Usando um teste para distinguir entre indivíduos com resultados normais e anormais, é importante entender como as características são distribuídas nas populações humanas.

A Figura 4–1 mostra a distribuição dos resultados do teste de tuberculina em uma população – o tamanho da enduração (área endurecida no local da injeção) em milímetros é mostrado no eixo horizontal e o número de indivíduos é indicado no eixo vertical. Um grande grupo obteve o valor de 0 mm – sem enduração – e outro grupo obteve aproximadamente 20 mm de enduração. Este tipo de distribuição, em que houve dois picos, é denominado *curva bimodal*. A distribuição bimodal permite a separação de indivíduos que não tiveram experiência anterior com a tuberculose (pessoas sem enduração, à esquerda) daquelas que tiveram uma experiência anterior com a tuberculose (aquelas com aproximadamente 20 mm de enduração, à direita). Embora alguns indivíduos permaneçam na "zona cinza", no centro, e possam

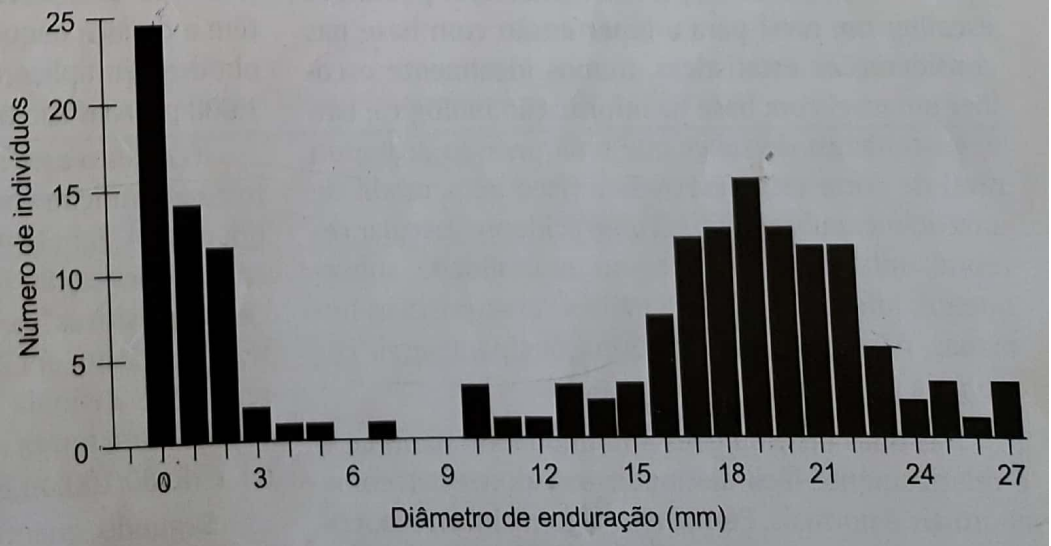


FIGURA 4–1. Distribuição das reações à tuberculina. (Adaptado de Edwards LB, Palmer CE, Magnus K: Vacina BCG: Estudos para o WHO Tuberculosis Research Office, Copenhagen, WHO Monograph Nº 12. WHO, Genebra, 1953.)

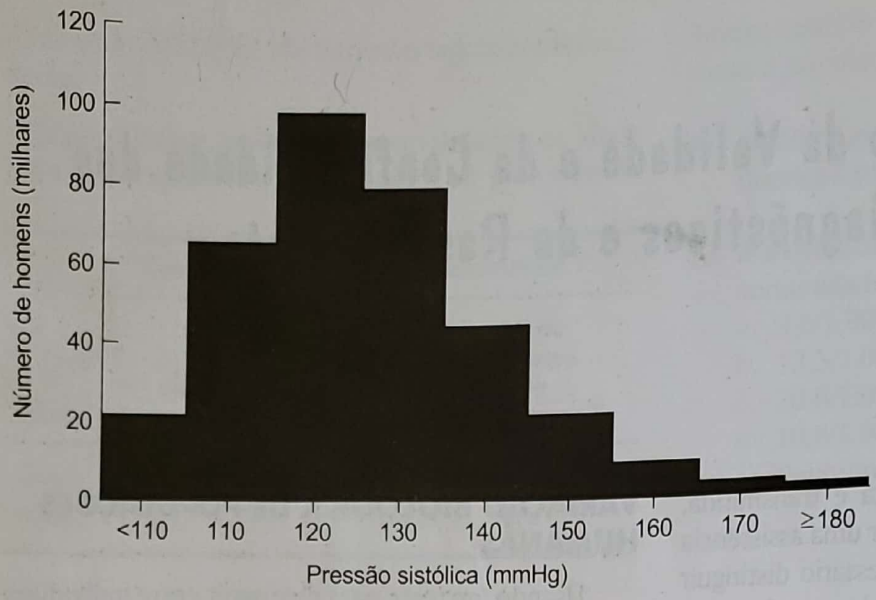


FIGURA 4–2. Distribuição da pressão sistólica para homens rastreados quanto à Pesquisa de Intervenção de Fatores Múltiplos de Risco. (Dados de Stamler J, Stamler R, Neaton JD: Pressão arterial, sistólica e diastólica, e os riscos cardiovasculares: Dados populacionais dos EUA. Arch Intern Med 153:598-615, 1993.)

pertencer a qualquer uma das curvas, a maioria da população pode ser facilmente distinguida usando as duas curvas. Então, quando uma característica apresenta uma distribuição bimodal, é relativamente fácil separar a maior parte da população em dois grupos (p. ex., doentes e saudáveis, apresentando certa condição ou anormalidade e não apresentando essa condição ou anormalidade).

Em geral, entretanto, a maioria das características humanas não é distribuída de modo bimodal. A Figura 4–2 mostra a distribuição das pressões sistólicas em um grupo particular. Nessa figura não há uma curva bimodal; o que vemos é uma *curva unimodal* – com um único pico. Assim, se queremos separar os hipertensos do grupo daqueles que não o são, o nível de corte da pressão arterial deve ser abaixo daquele cujas pessoas são denominadas hipertensas e abaixo daquele em que são designadas como normotensas. Nenhum nível óbvio de pressão sangüínea distingue os normotensos dos hipertensos. Embora possamos escolher um nível para a hipertensão com base nas considerações estatísticas, iremos idealmente escolher um nível com base na informação biológica; isto é, gostaríamos de saber que uma pressão acima do nível de corte está associada a risco aumentado de uma doença subseqüente, como acidente vascular cerebral, infarto miocárdico, ou mortalidade subseqüente. Infelizmente, para muitas características humanas, não temos essas informações para servir como guia para estabelecer esse nível.

Nas duas distribuições – unimodal ou bimodal – é relativamente fácil distinguir os valores extremos anormais e normais. Permanece alguma incerteza, entretanto, em casos que estão na zona cinza dos dois tipos de curva.

VALIDADE DOS TESTES DE RASTREAMENTO

A *validade* de um teste é determinada como a habilidade de um teste de distinguir quem tem a doença daqueles que não a têm. A validade possui dois componentes: sensibilidade e especificidade. A *sensibilidade* do teste é definida como a habilidade deste em identificar corretamente aqueles que têm a doença. A *especificidade* do teste é definida como a habilidade deste em identificar corretamente aqueles que *não têm* a doença.

Testes com Resultados Dicotomizados (Positivos ou Negativos)

Vamos supor que temos uma população hipotética de 1.000 pessoas e que 100 pessoas têm uma certa doença e 900 não. Um teste está disponível e pode apresentar resultados positivos ou negativos. Queremos usar esse teste para tentar separar pessoas que têm a doença daquelas que não a têm. Os resultados obtidos pela aplicação do teste para esta população de 1.000 pessoas são mostrados na Tabela 4–1.

O quanto este teste é bom? Primeiro, o quanto o teste identificou corretamente aqueles que tinham a doença? A Tabela 4–1 indica que, das 100 pessoas com a doença, 80 foram corretamente identificadas como "positivas" pelo teste e uma identificação positiva foi ausente em 20. Portanto, a *sensibilidade* do teste, que é definida como a proporção de pessoas doentes que foram corretamente identificadas como tal, é de 80/100 ou 80%.

Segundo, quanto o teste identificou corretamente aqueles que não tinham a doença? Observando novamente a Tabela 4–1, de 900 pessoas que não ti-

TABELA 4–1. *Conceito de Sensibilidade e Especificidade das Pesquisas por Rastreamento*

Exemplo: considere uma população de 1.000 pessoas, onde 100 apresentavam a doença e 900 não apresentavam a doença

Teste de rastreamento para identificar as 100 pessoas com a doença

	Características Verdadeiras da População		
Resultado do Rastreamento	**Doença**	**Sem Doença**	**Total**
Positivo	80	100	180
Negativo	20	800	820
Total	100	900	1.000

$$\text{Sensibilidade} = \frac{80}{100} = 80\%$$

$$\text{Especificidade} = \frac{800}{900} = 89\%$$

nham a doença, o teste identificou 800 como "negativas". A *epecificidade* do teste, que é definida como a proporção de pessoas não doentes que foram corretamente identificadas como negativas pelo teste, foi de 800/900 ou 89%.

Observe que, para calcular a sensibilidade e a especificidade de um teste, precisamos saber quem "realmente" apresenta a doença e quem não apresenta a doença de outra fonte, diferente do teste que estamos usando. Estamos, na verdade, comparando o resultado de nosso teste com algum "padrão-ouro" – uma fonte externa de "verdade" a respeito do estado da doença de cada indivíduo na população. Às vezes, essa verdade pode ser o resultado de um outro teste que foi utilizado e algumas vezes é o resultado de um teste mais definitivo e geralmente mais invasivo (p. ex., cateterização cardíaca ou biópsia tecidual). Entretanto, na vida real, quando utilizamos um teste para identificar pessoas doentes e não doentes, claramente não sabemos quem tem a doença e quem não tem a doença. (Se isso já estivesse estabelecido, o teste seria inútil.) Mas para avaliar quantitativamente a sensibilidade e especificidade de um teste, devemos ter uma outra fonte de verdade para podermos comparar os resultados.

A Tabela 4–2 compara os resultados dos testes dicotomizados (com resultados tanto positivos como negativos) com o estado atual da doença. Idealmente, gostaríamos que todos os elementos testados caíssem em uma das duas células mostradas acima, à esquerda, e abaixo, à direita, na tabela: as pessoas com a doença que são corretamente chamadas "positivas" pelo teste *(positivos verdadeiros)* e as pessoas sem a doença, denominadas corretamente "negativas" pelo teste *(negativos verdadeiros)*. Infelizmente, isso é raro, quando for o caso. Algumas pessoas que não apresentam a doença são chamadas de "positivas" pelo teste *(falsos positivos)*, e algumas pessoas com a doença são erroneamente chamadas "negativas" *(falsos negativos)*.

Por que essas questões são importantes? Quando conduzimos um programa de rastreamento, geralmente temos um grande grupo de pessoas que são consideradas positivas, incluindo as pessoas que realmente têm a doença (verdadeiros positivos) e as pessoas que não têm a doença (falsos positivos). A questão dos *falsos positivos* é importante porque todas as pessoas que foram rastreadas como positivas são incluídas na realização de testes mais sofisticados e mais caros. Dos vários problemas que resultam, o primeiro é a sobrecarga no sistema de atendimento à

TABELA 4-2. *Comparação dos Resultados de um Teste Dicotomizado com o Estado da Doença*

	População	
Resultados do Teste	**Com a Doença**	**Sem a Doença**
Positivo	Positivo verdadeiro (PV) = Possui a doença e um teste positivo	Falso positivo (FP) = sem doença, mas com teste positivo
Negativo	Falso negativo (FN) = Possui a doença, mas tem um teste negativo	Negativo verdadeiro (NV) = Sem a doença e com teste negativo

$$\text{Sensibilidade} = \frac{PV}{PV + FN} \qquad \text{Especificidade} = \frac{NV}{NV + FP}$$

saúde. Outros são ansiedade e preocupação, provocadas nas pessoas que souberam que apresentavam o teste positivo. Uma considerável evidência indica que muitas pessoas, rotuladas "positivas" por um teste de rastreamento, nunca tiveram esse rótulo completamente apagado, mesmo que os resultados de avaliações subseqüentes fossem negativos. Por exemplo, crianças rotuladas como "positivas" em um programa de rastreamento para doença cardíaca foram tratadas como deficientes pelos pais e pelas pessoas da escola, mesmo após testes negativos. Além disso, tais indivíduos podem ser barrados em relação a um emprego ou seguro por uma interpretação errônea de um resultado positivo de um teste de rastreamento, mesmo que os testes posteriores não encontrem qualquer achado positivo.

Por que o problema dos *falsos negativos* é importante? Se uma pessoa tem a doença, mas é erroneamente informada que seu resultado é negativo, e se a doença é séria e uma intervenção efetiva está disponível, o problema é mais sério. Por exemplo, se a doença é um tipo de câncer curável somente nos estágios iniciais, um resultado falso negativo poderia representar uma sentença de morte. Portanto, a importância dos resultados falsos negativos depende da natureza e gravidade da doença rastreada, da eficiência de medidas de intervenção disponíveis, e se a eficiência é maior se a intervenção for realizada no início da história natural da doença.

Testes de Variáveis Contínuas

Até agora discutimos um teste com somente dois resultados possíveis: positivo ou negativo. Mas nós sempre testamos uma variável contínua, como a pressão arterial ou nível de glicose sangüínea, em que não há um resultado "positivo" ou "negativo". Uma decisão deve ser tomada, no estabelecimento de um nível de corte acima do qual o resultado é considerado positivo e, abaixo, é considerado negativo. Vamos considerar os diagramas mostrados na Figura 4–3.

A Figura 4–3A mostra uma população de 20 diabéticos e 20 não diabéticos que estão sendo rastreados utilizando o teste de glicemia, que é mostrado ao longo do eixo vertical, do mais alto para o mais baixo. Os diabéticos são representados por círculos sólidos e os não diabéticos por círculos sombreados. Observamos que, embora os níveis de açúcar no sangue tendam a ser maiores nos diabéticos do que nos não diabéticos, nenhum nível claramente separa os dois grupos; há alguma sobreposição de diabéticos e não diabéticos em cada nível sangüíneo. Contudo, devemos selecionar um ponto de corte para que os resultados que estejam acima do corte possam ser chamados de "positivos" e os pacientes chamados para novos testes, daqueles em que os resultados estejam abaixo do ponto e chamados "negativos", e não sejam chamados para outros testes.

Vamos supor que um ponto de corte relativamente alto seja escolhido (Fig. 4–3B). Claramente, muitos dos diabéticos não serão identificados como positivos; de outro modo, a maioria dos não diabéticos será corretamente identificada como negativa. Se esses resultados forem distribuídos em uma tabela de 2×2, a sensibilidade do teste que utiliza este ponto de corte será de 25% (5/20) e a especificidade será de 90% (18/20).

O que ocorre se um ponto de corte baixo for escolhido (Fig. 4–3C)? Muito poucos diabéticos serão diagnosticados erroneamente. Qual é o problema então? Uma grande proporção de não diabéticos agora será identificada como positiva pelo teste. Como foi visto na tabela 2×2, a sensibilidade é agora de 85% (17/20), mas a especificidade é somente de 30% (6/20).

A dificuldade que existe no mundo real é que nenhuma linha vertical separa os diabéticos dos não diabéticos, e que eles são, de fato, misturados (Fig. 4–3D); de fato, eles não são perceptíveis pelos círculos sólidos ou sombreados (Fig. 4–3E). Então, se um nível alto de corte for utilizado (Fig. 4–3F), todos aqueles com resultados abaixo da linha serão informados que não têm a doença e não serão mais avaliados; se o ponto de corte for baixo (Fig. 4–3G), todos aqueles com resultados acima da linha serão avaliados com mais testes.

A Figura 4–4 mostra dados reais, relacionados à distribuição de níveis de glicose sangüínea em diabéticos e não diabéticos. Vamos supor que iremos rastrear essa população. Se decidirmos fixar o nível de corte para identificar todos os diabéticos (100% de sensibilidade), poderemos fixar o nível para 80 mg/dl. O problema é que, entretanto, fazendo isso, iremos também determinar muitos não diabéticos como positivos (especificidade muito baixa). Por outro lado, se fixarmos o nível em 200 mg/dl, então determinaremos todos os não diabéticos como negativos (100% de especificidade), agora iremos perder muitos dos diabéticos verdadeiros (sensibilidade muito baixa). Dessa forma, há uma troca entre a sensibilidade e a especificidade: se aumentamos a sensibilidade, abaixando o nível de corte, diminuímos a especificidade; se aumentamos a especificidade aumentando o nível de corte, diminuímos a sensibilidade. Para citar um sábio desconhecido: "Não há nada como um almoço grátis".

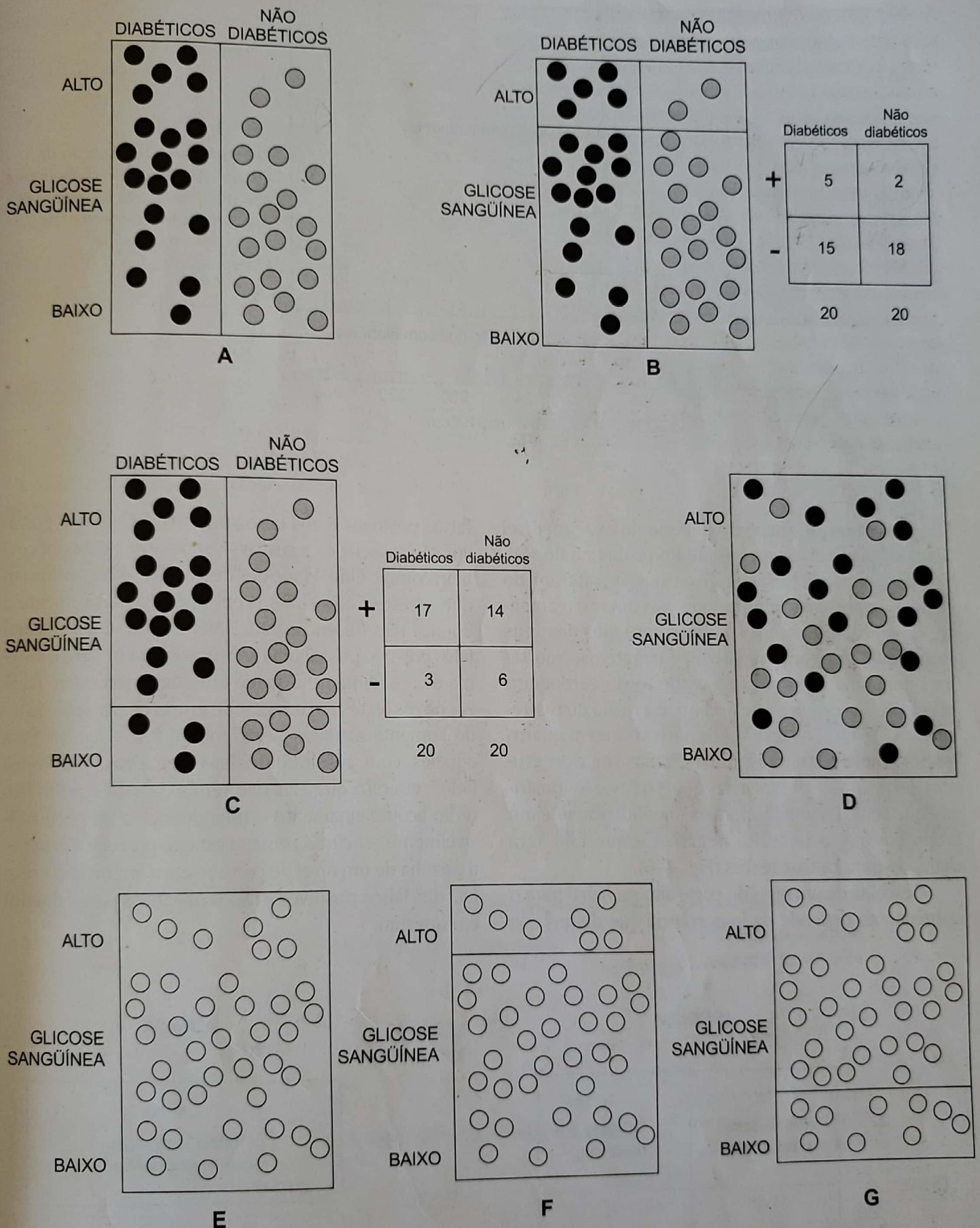


FIGURA 4–3. A-G, Rastreamento para o diabetes em uma população hipotética com a prevalência de 50%. Efeitos da escolha de diferentes níveis de corte para um teste positivo. (Veja o texto.)

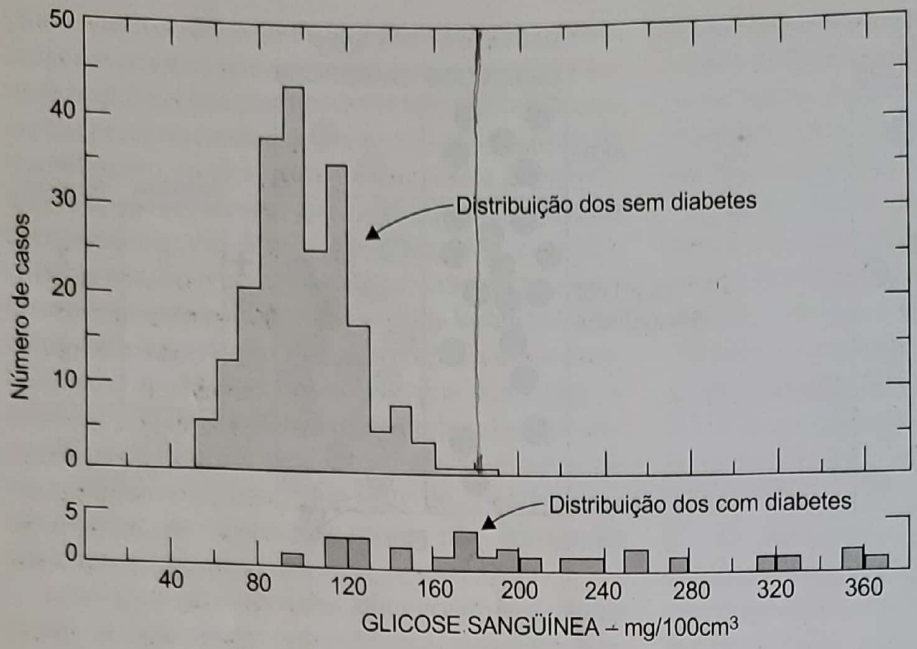


FIGURA 4–4. Distribuição da glicose sangüínea em diabéticos e não diabéticos. (De Blumberg M: Avaliando procedimentos de rastreamento de saúde. Operations Res 5:351-360, 1957.)

O dilema que envolve a decisão de fixar um nível alto ou baixo de corte consiste no problema dos falsos positivos e falsos negativos que resultaram do teste. É importante lembrar que, em um rastreamento, classificamos grupos somente baseados nos seus resultados, tais como positivos e negativos. Não temos nenhuma informação a respeito do verdadeiro estado de sua doença que, é claro, é a razão do rastreamento. De fato, os resultados não produzem quatro grupos, como visto na Figura 4–5, mas sim dois grupos: um grupo de pessoas nas quais o teste foi positivo e que serão chamadas para exames adicionais e um grupo em que o teste foi negativo e que não será chamado para outros testes (Fig. 4–6).

A escolha de um nível de corte alto ou baixo para o rastreamento depende da importância que damos aos falsos positivos e aos falsos negativos. Os falsos positivos são associados a custos – emocional e econômico – bem como a dificuldade de "desrotular" uma pessoa em que o teste foi positivo e em quem posteriormente a doença não foi encontrada. Além disso, um resultado falso positivo possui uma sobrecarga maior para o sistema de atendimento à saúde em que é necessário realizar novos testes em um grande grupo de pessoas, quando somente algumas delas podem ter a doença. Para aqueles com resultados falsos negativos, por outro lado, será dito que eles não têm a doença e assim não terão acompanhamento e uma doença séria pode possivelmente ser curada em um estágio precoce. Portanto, a escolha de um nível de corte relata a importância relativa dos falsos positivos e falsos negativos para a doença em questão.

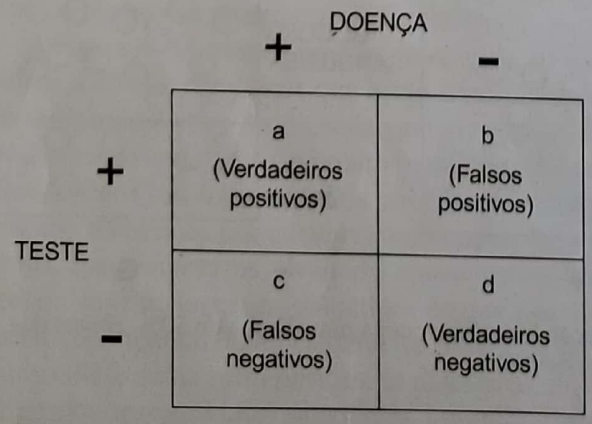


FIGURA 4–5. Diagrama indicando quatro possíveis grupos resultantes de testes de rastreamento, utilizando-se um teste dicotomizado.

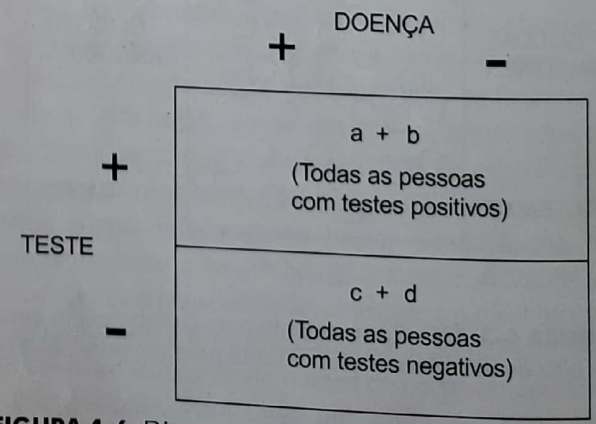


FIGURA 4–6. Diagrama agrupando todas as pessoas com testes positivos e todas as pessoas com testes negativos no rastreamento.

Considere: Prevalência da doença = 5%
População = 10.000

TESTE 1 (Glicose sangüínea) Sensibilidade = 70%
Especificidade = 80%

RESULTADOS DOS TESTES	+ DIABETES	-	
+	350	1.900	2.250
-	150	7.600	7.750
	500	9.500	10.000

FIGURA 4–7. Exemplo hipotético de programa de rastreamento em dois estágios: I.

USO DE MÚLTIPLOS TESTES

Geralmente podem ser utilizados vários testes de rastreamento – seqüencial ou simultaneamente. Os resultados dessas abordagens são descritos resumidamente.

Teste Seqüencial (Dois Estágios)

Em testes seqüenciais ou rastreamento em dois estágios, um teste mais barato, menos invasivo e mais confortável é geralmente realizado primeiramente, e aqueles com rastreamento positivo são reconvocados para mais testes, mais caros, mais invasivos e mais desconfortáveis, que tenham sensibilidade e especificidade maiores. Espera-se que, fazendo mais testes somente naqueles em que o rastreamento tenha sido positivo, o problema do falso positivo seja reduzido.

Considere o exemplo hipotético na Figura 4–7, em que uma população é rastreada para o diabetes utilizando um teste com sensibilidade de 70% e especificidade de 80%. Como estes dados mostrados na tabela foram obtidos? A prevalência da doença nessa população foi dada como sendo de 5%, de maneira que, na população de 10.000, 500 pessoas têm a doença. Com a sensibilidade de 70%, o teste identificará corretamente 350 das 500 pessoas que têm a doença. Com a especificidade de 80%, o teste identificará corretamente como não diabéticos 7.600 das 9.500 pessoas que não apresentam diabetes; entretanto, 1.900 destas 9.500 terão resultados positivos. Portanto, um total de 2.250 pessoas terão um teste positivo e serão trazidas para um segundo teste. (Lembre-se que, na vida real, não temos a linha vertical, separando os diabéticos dos não diabéticos, e que não sabemos que 350 dos 2.250 têm o diabetes.)

Agora, aquelas 2.250 pessoas são trazidas para a realização de rastreamento, usando um segundo teste (tal como o teste de tolerância à glicose), que neste exemplo é considerado como tendo uma sensibilidade de 90% e uma especificidade de 90%. A Figura 4–8 novamente mostra o teste 1 com o teste 2, que lida somente com as 2.250 pessoas que foram positivas no primeiro teste e foram trazidas para a realização de rastreamento em segundo estágio.

Como as 350 pessoas (de 2.250) têm a doença e o teste tem uma sensibilidade de 90%, 315 daquelas 350 serão corretamente identificadas como positivas. Como 1.900 (de 2.250) não apresentam o diabetes e o teste de especificidade é de 90%, 1.710 das 1.900

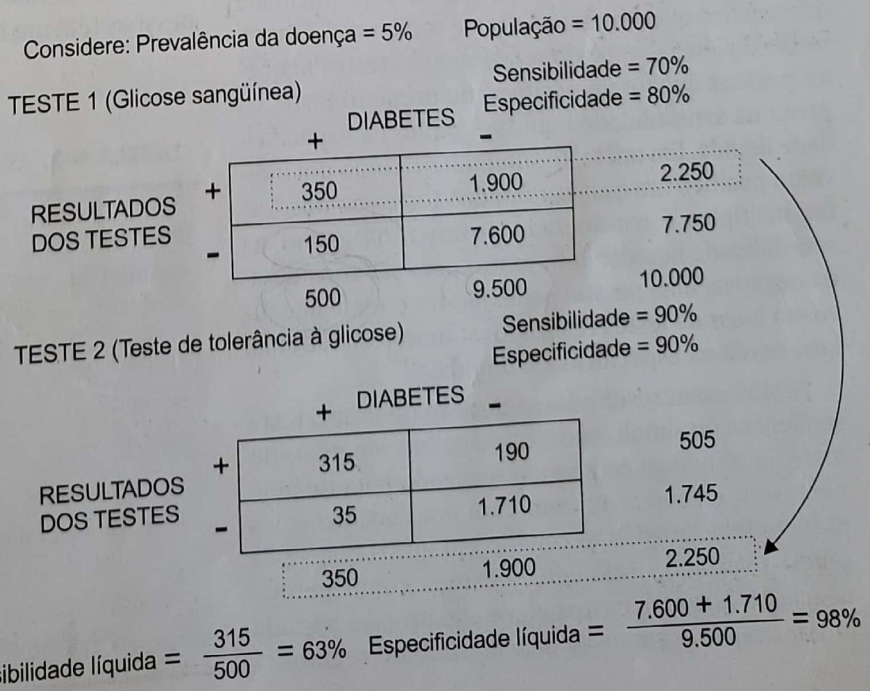


FIGURA 4–8. Exemplo hipotético de programa de rastreamento em dois estágios: II.

serão corretamente identificadas como negativas e 190 serão falsos positivos.

Agora somos capazes de calcular a *sensibilidade líquida* e a *especificidade líquida*, utilizando ambos os testes em seqüência. Após o término de ambos os testes, 315 pessoas do total de 500 nesta população de 10.000 foram corretamente determinadas como positivas: 315/500 = 63% *sensibilidade líquida*. Então, há uma perda na sensibilidade quando utilizamos ambos os testes. Para calcular a *especificidade líquida*, observe que 7.600 pessoas de 9.500 na população que não tinham o diabetes foram corretamente determinadas como negativas no rastreamento de primeiro estágio e não foram mais testadas; outros 1.710 dos 9.500 não diabéticos foram corretamente determinadas como negativas no rastreamento de segundo estágio. Então um total de 7.600 + 1.710 dos 9.500 não diabéticos foram corretamente determinados como negativos. 9.310/9.500 = 98% *especificidade líquida*. Portanto, o uso de ambos os testes resultou em ganho na *especificidade líquida*.

Teste Simultâneo

Em um estudo clínico, geralmente são utilizados múltiplos testes simultaneamente. Por exemplo, um paciente admitido em um hospital pode ter uma quantidade de testes realizados no momento de admissão. Quando são utilizados múltiplos testes simultaneamente para detectar uma doença específica, o indivíduo é geralmente considerado "positivo" se ele tiver um resultado positivo em um ou mais testes. O indivíduo é considerado "negativo" se os testes foram todos negativos. Os efeitos de tal abordagem de testes com base na sensibilidade e especificidade diferem dos que resultam do teste seqüencial. Em um teste seqüencial, quando refazemos os testes naquelas pessoas que foram positivas no primeiro teste, há perda na sensibilidade líquida e ganho na especificidade líquida. Em um teste simultâneo, como um indivíduo positivo em *qualquer* um dos testes ou nos testes múltiplos é considerado positivo, há ganho na sensibilidade líquida. Entretanto, para ser considerada negativa, uma pessoa pode não ter o teste negativo em *todos* os testes realizados. Como resultado, há uma perda na especificidade líquida.

Dado esses resultados, a decisão de usar o teste seqüencial ou simultâneo é baseada nos objetivos do teste, incluindo se está sendo realizado para rastreamento ou diagnóstico, bem com em considerações práticas relacionadas ao cenário em que o teste está sendo realizado, incluindo questões como permanência no hospital, custos e grau de invasão de cada um dos testes e extensão da cobertura de seguros.

VALOR PREDITIVO DE UM TESTE

Até agora, perguntamos: o quanto o teste é bom para identificar as pessoas com a doença e as pessoas sem a doença? Este é um assunto importante, particularmente no rastreamento de populações livres. De fato, estamos perguntando: se rastrearmos uma população, que proporção de pessoas que possuem uma doença será corretamente identificada? Essa é claramente uma consideração importante de saúde pública. Na clínica, entretanto, uma questão diferente pode ser importante para o médico: se o teste resulta em positivo neste paciente, qual a probabilidade de que esse paciente tenha a doença? Isso é chamado de *valor preditivo positivo* do teste. Em outras palavras, que proporção de pacientes com teste positivo atualmente possui a doença em questão? Para calcular o valor preditivo, dividimos o número de verdadeiros positivos pelo número total de elementos positivos (verdadeiros + falsos positivos).

Vamos retornar ao exemplo apresentado na Tabela 4–1, em que uma população de 1.000 pessoas é rastreada. Como visto na Tabela 4–3, uma tabela 2 × 2 mostra o resultado de um rastreamento dicotomizado na população. De 1.000 elementos, 180 têm um teste positivo; destes 180, 80 têm a doença. O *valor preditivo positivo* é então 80/180 ou 44%.

Uma pergunta paralela pode ser feita em relação aos testes negativos: se o resultado do teste é negativo, qual a probabilidade desse paciente não ter a doença? Isso é chamado de *valor preditivo negativo* do teste. Ele é calculado dividindo-se o número dos verdadeiro negativos por todos aqueles que foram negativos (verdadeiros negativos + falsos negativos). Observando novamente o exemplo da Tabela 4–3, 820 pessoas têm um resultado negativo e, destas, 800 não

TABELA 4–3. *Valor Preditivo de um Teste*

Resultados do Teste	População		
	Doença	Sem Doença	Total
Positivo	80	100	180
Negativo	20	800	820
Total	100	900	1.000

$$\text{Valor preditivo positivo} = \frac{80}{180} = 44\%$$

$$\text{Valor preditivo negativo} = \frac{800}{820} = 98\%$$

TABELA 4–4. *Relação da Prevalência da Doença e o Valor Preditivo*

	Exemplo: sensibilidade = 99%, especificidade = 95%				
Prevalência da Doença	**Resultados do Teste**	**Doente**	**Não-Doente**	**Totais**	**Valor Preditivo**
1%	+	99	495	594	$\frac{99}{594} = 17\%$
	–	1	9.405	9.406	
	Totais	100	9.900	10.000	
5%	+	495	475	970	$\frac{495}{970} = 51\%$
	–	5	9.025	9.303	
	Totais	500	9.500	10.000	

têm a doença. Portanto, o *valor preditivo negativo* é 800/820 ou 98%. Na discussão do valor preditivo que segue, o termo *valor preditivo* é usado para denotar o valor preditivo positivo do teste.

Cada teste que um médico realiza – histórico, exame físico, testes laboratoriais, raios X, eletrocardiograma e outros procedimentos – é utilizado para reforçar a habilidade do médico em fazer um diagnóstico correto. O que ele ou ela quer saber após aplicar um teste ao paciente é: se o resultado for positivo, qual a probabilidade de o paciente ter a doença?

Diferente da sensibilidade e da especificidade do teste, que podem ser consideradas características do teste utilizado, o valor preditivo é afetado por dois fatores: a prevalência da doença na população testada e, quando a doença é infreqüente, a especificidade do teste utilizado. Ambas as relações são discutidas nas seções a seguir.

Relação de um Valor Preditivo com a Prevalência da Doença

A relação entre o valor preditivo e a prevalência da doença pode ser vista no exemplo dado na Tabela 4–4.

Primeiro, vamos direcionar nossa atenção para a parte superior da tabela. Considere que estamos utilizando um teste com uma sensibilidade de 99% e uma especificidade de 95% em uma população de 10.000 pessoas em que a prevalência da doença seja de 1%. Como a prevalência é de 1%, 100 das 10.000 pessoas apresentam a doença e 9.900 não apresentam. Com a sensibilidade de 99%, o teste identificará corretamente 99 entre 100 pessoas que têm a doença. Com a sensibilidade de 95%, o teste identificará corretamente 9.405 das 9.900 pessoas que não têm a doença.

Dessa forma, nesta população com 1% de prevalência, 594 pessoas são chamadas positivas pelo teste (99 + 495).

Entretanto, destas 594, 495 (83%) são falsos positivos e o valor preditivo positivo é, portanto, 99/594 ou somente 17%.

Vamos agora aplicar o mesmo teste – com a mesma sensibilidade e especificidade – em uma população com uma prevalência mais alta para a doença, 5%, como visto na parte inferior da Tabela 4–4. Usando cálculos similares àqueles utilizados na parte superior da tabela, o valor preditivo positivo é agora de 51%. Então, uma prevalência maior na população rastreada leva a aumento no valor preditivo positivo, utilizando-se o mesmo teste.

A Figura 4–9 mostra a relação entre a prevalência da doença e o valor preditivo: claramente, a maior parte do ganho no valor preditivo ocorre com o

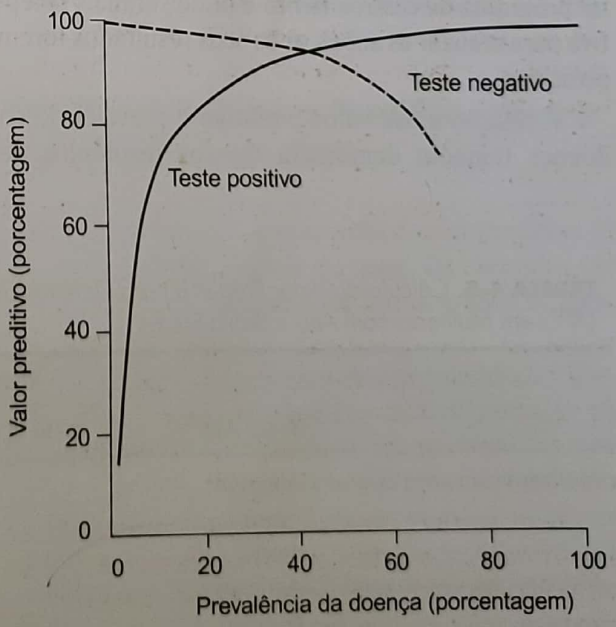


FIGURA 4–9. Relação entre a prevalência da doença e o valor preditivo em um teste com 95% de sensibilidade e 95% de especificidade. (De Mausner JS, Kramer S: Mausner and Bahn Epidemiology: An Introductory Text. Philadelphia, WB Saunders, 1985, p 221.)

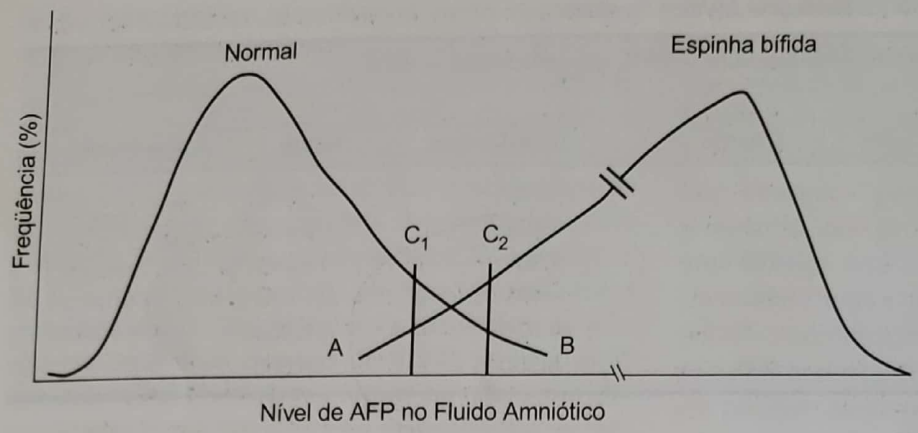


FIGURA 4–10. Níveis de alfafetoproteína (AFP) no fluido amniótico em pacientes normais e com espinha bífida. (De Sheffield LJ, Sackett DL, Goldsmith CH, *et al.*: Abordagem clínica do uso de valores preditivos no diagnóstico pré-natal de defeitos do tubo neural. Am J Obstet Gynecol 145:319-324, 1983.)

aumento na prevalência das taxas inferiores da prevalência da doença.

Por que devemos ter cuidado com a relação entre valor preditivo e prevalência da doença? Vimos que, quanto maior a prevalência, maior será o valor preditivo. Portanto, um programa de rastreamento é mais produtivo e eficiente se for direcionado a uma população-alvo de alto risco. O rastreamento de uma população inteira quanto a uma doença relativamente infreqüente pode ser muito dispendioso e pode ter poucos casos detectados para a quantidade de esforços envolvidos. Entretanto, se um subgrupo de alto risco puder ser identificado e o rastreamento puder ser direcionado a ele, o programa provavelmente será muito mais produtivo. Além do mais, uma população de alto risco pode ser mais motivada a participar de tal programa de rastreamento e pode ser mais receptiva para realizar as ações se os seus resultados forem positivos.

A relação entre valor preditivo e prevalência da doença também demonstra que os resultados de qualquer teste devem ser interpretados no contexto da prevalência da doença na população em que se originou. Um exemplo interessante é observado com o uso da determinação da alfafetoproteína (AFP) no fluido amniótico para o diagnóstico pré-natal de espinha bífida. A Figura 4–10 mostra a distribuição dos níveis de AFP no fluido amniótico em gestações normais e em gestações em que o feto apresentava a espinha bífida, um defeito do tubo neural. Embora a distribuição seja bimodal, existe uma extensão em que as curvas se sobrepõem e, nessa série, pode não ser sempre claro qual é a curva da mãe e qual a curva da criança. Sheffield *et al.*[1] revisaram a literatura e construíram populações artificiais de 10.000 mulheres rastreadas quanto a AFP no líquido amniótico para identificarem fetos com espinha bífida. Eles criaram duas populações: uma com alto risco para espinha bífida e outra com risco normal.

A Tabela 4–5 mostra os cálculos para as mulheres de alto e baixo riscos. Quais as mulheres que possuem um alto risco de ter uma criança com espinha

TABELA 4–5. *Cálculo do Valor Preditivo para Defeitos do Tubo Neural (DTN)* para o Teste de Alfafetoproteína (AFP) em Mulheres com Alto e Baixo Riscos*

		Resultado da Gravidez			
	Teste AFP	**DTN**	**Normal**	**100.000**	**Valor Preditivo (%)**
Mulheres com alto risco	Anormal	87	18	105	82,9
	Normal	13	9.882	9.895	99,9
	Total	100	9.900	10.000	
Mulheres com baixo risco	Anormal	128	179	307	41,7
	Normal	19	99.674	99.693	99,98
	Total	147	99.853	100.000	

*Espinha bífida ou encefalocele.
De Sheffield LJ, Sackett DL, Goldsmith CH, *et al.*: Uma abordagem clínica para o uso dos valores preditivos no diagnóstico pré-natal dos defeitos do tubo neural. Am J Obstet Gynecol 145:319-324, 1983.

bífida? É sabido que a mulher que teve uma criança com um defeito do tubo neural possui risco aumentado porque o defeito tende a se repetir nos irmãos. Nestes cálculos, o valor preditivo positivo encontrado é de 82,9%. Quais as mulheres com baixo risco, mas que ainda devem realizar amniocentese? Estas são as mulheres mais velhas que fazem amniocentese pela possibilidade de o feto apresentar síndrome de Down ou outro defeito associado à gestação em idade avançada. O risco de espinha bífida, entretanto, não é relacionado com a idade materna, portanto essas mulheres não possuem risco aumentado. O cálculo mostra que, usando-se o mesmo teste para a AFP, como foi usado para as mulheres de alto risco, o valor preditivo positivo foi de somente 41,7%, consideravelmente menor do que o do grupo de alto risco.

Portanto, vemos que o mesmo teste pode apresentar um valor preditivo diferente quando é administrado a uma população de alto risco (com alta prevalência) ou a uma população de baixo risco (baixa prevalência). Isso possui implicações clínicas importantes: uma mulher pode decidir terminar a gestação e o médico pode formular um conselho para essa mulher, baseado nos resultados dos testes. Mas o mesmo resultado pode ser interpretado diferentemente, dependendo se a mulher é proveniente de um *pool* de alto risco ou de baixo risco, em que será refletido o valor preditivo positivo do teste. Conseqüentemente, o teste por si mesmo não é suficiente para servir como guia, sem levar em conta outras considerações acima descritas.

Os exemplos reais a seguir demonstram a importância deste capítulo:

> *O chefe dos bombeiros consultou um cardiologista universitário porque o médico do departamento de bombeiros lera um artigo em uma reportagem médica de que um certo achado eletrocardiográfico era um fator preditivo alto de doença coronariana séria, geralmente não reconhecida. Baseado nesse jornal, o médico do departamento dos bombeiros estava desqualificando muitos jovens capazes para a atividade. O cardiologista leu o artigo e achou que o estudo fora realizado em pacientes hospitalizados.*

Qual o problema? Como os pacientes hospitalizados apresentam uma prevalência muito maior de doença cardíaca do que o grupo de jovens bombeiros, o médico do departamento dos bombeiros utilizou erroneamente o alto valor preditivo obtido no estudo de uma população com alta prevalência e, inapropriadamente, aplicou a uma população de baixa prevalência de bombeiros saudáveis, em que o mesmo teste poderia apresentar um valor preditivo muito menor.

Outro exemplo:

> *Um médico visitou seu clínico geral para exame médico anual, que incluía pesquisa de sangue oculto nas fezes. Uma das três amostras examinadas no teste foi positiva. O clínico contou ao seu paciente que o resultado não tinha nenhum significado porque ele regularmente encontrava muitos testes falsos positivos na sua prática clínica. O teste foi repetido e agora todas as três amostras foram negativas. Mesmo assim, o clínico encaminhou seu paciente para um gastroenterologista. O gastroenterologista disse: "Em minha experiência, um achado positivo é sério. Tal achado é freqüentemente associado a patologias gastrintestinais. Os testes negativos subseqüentes não significam nada, porque você poderia ter um tumor que somente sangra intermitentemente."*

Quem estava correto nesse episódio? A resposta é que ambos estavam corretos. O clínico deu sua avaliação do valor preditivo baseado em sua experiência na sua prática médica – uma população com baixa prevalência de doenças gastrintestinais sérias. O gastroenterologista, por outro lado, deu sua avaliação do valor preditivo do teste baseado em sua experiência em sua prática – a prática em que a maioria dos pacientes apresenta doença gastrintestinal séria – uma população de alta prevalência.

Relação dos Valores Preditivos para a Especificidade do Teste

Um segundo fator que afeta o valor preditivo de um teste é a *especificidade* do teste. Os exemplos são mostrados primeiramente de forma gráfica e, após, de forma tabulada.

A Figura 4–11 A a D, diagrama os resultados de rastreamento populacional; entretanto as tabelas 2 × 2 nesses valores são diferentes dos números anteriores: o tamanho de cada célula é proporcional à população que representa. Em cada número, as células que representam pessoas com teste positivo são sombreadas em cinza, sendo que estas são as células que serão utilizadas no cálculo do valor preditivo positivo.

A Figura 4–11A apresenta a população de base rastreada, que é utilizada em nossa discussão: uma população de 1.000 pessoas, em que a prevalência é

de 50%; então, 500 pessoas possuem a doença e 500 não. Analisando esse número, consideramos que o teste de rastreamento utilizado tem uma sensibilidade de 50% e uma especificidade de 50%. Como as 500 pessoas testadas foram positivas e outras 250 destas têm a doença, o valor preditivo é de 250/500 ou 50%.

Felizmente, a prevalência da maioria das doenças é muito menor que 50%; geralmente lidamos com doenças relativamente infreqüentes. A Figura 4–11B, então, considera uma prevalência menor, 20% (embora seja uma prevalência excepcionalmente alta para a maioria das doenças); a sensibilidade e a especificidade permanecem em 50%. Agora somente 200, das 1.000 pessoas, apresentam a doença e a linha vertical separando as pessoas doentes das não doentes é deslocada para a esquerda. O valor preditivo é agora calculado como 100/500 ou 20%.

Visto que estamos rastreando uma população com uma taxa de prevalência mais baixa, podemos melhorar o valor preditivo? Qual seria o efeito no valor

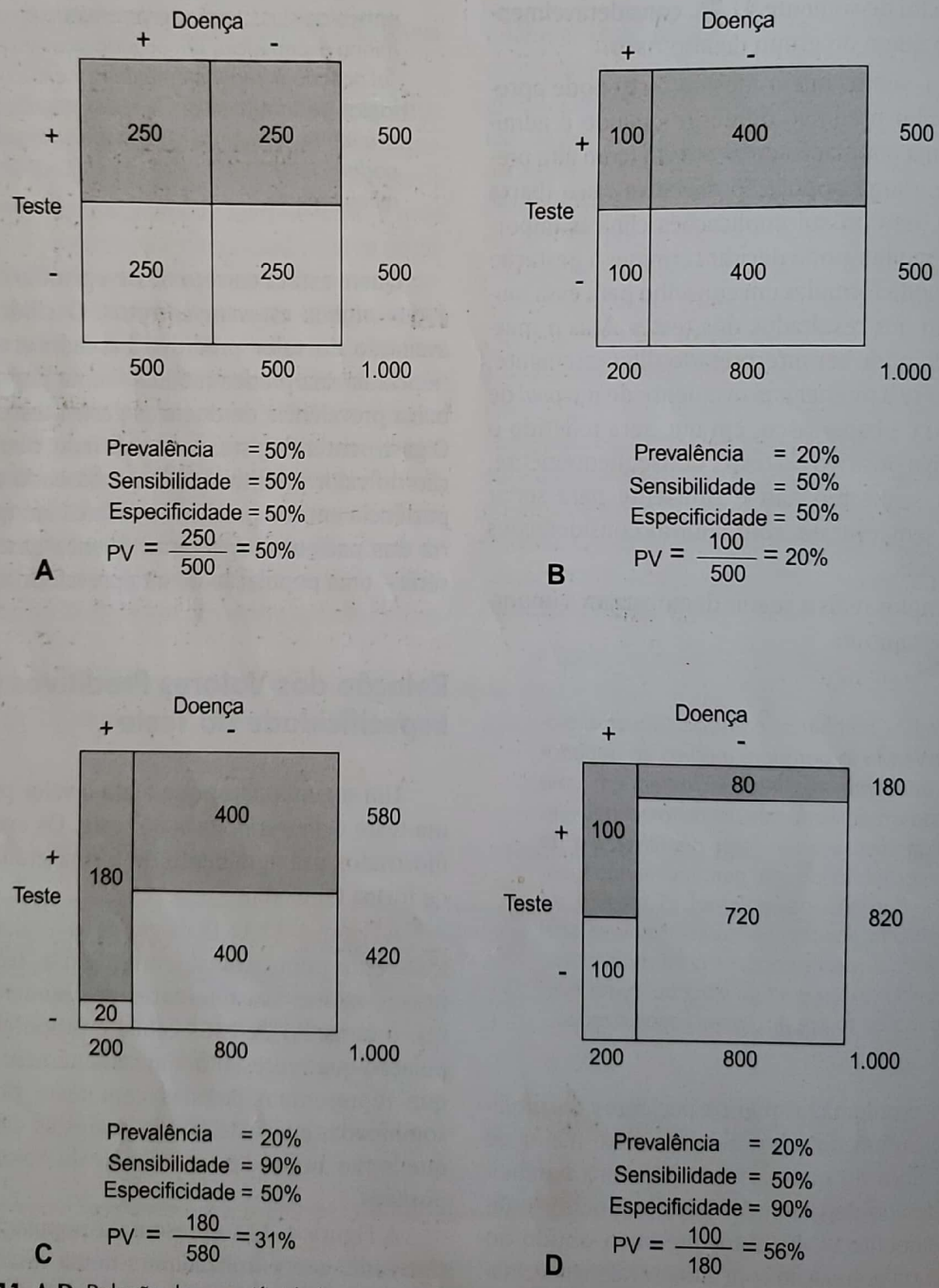


FIGURA 4–11. A-D, Relação da especificidade com os valores preditivos.

TABELA 4–6. *Relação da Especificidade e o Valor Preditivo*

Exemplo: prevalência = 10%, sensibilidade = 100%					
Especificidade	**Resultados do Teste**	**Doente**	**Não-Doente**	**Totais**	**Valor Preditivo**
70%	+	1.000	2.700	3.700	$\frac{1.000}{3.700} = 27\%$
	–	0	6.300	6.300	
	Totais	1.000	9.000	10.000	
95%	+	1.000	450	1.450	$\frac{1.000}{1.450} = 69\%$
	–	0	8.550	8.550	
	Totais	1.000	9.000	10.000	

preditivo se aumentássemos a sensibilidade para o teste? A Figura 4–11C mostra o resultado quando deixamos a prevalência em 20% e a especificidade em 50% mas aumentamos a sensibilidade para 90%. O valor preditivo é agora de 180/580 ou 31%, um aumento modesto.

E se em vez de aumentar a sensibilidade do teste, aumentássemos a especificidade? A Figura 4–11D mostra o resultado quando a prevalência permanece 20% e a sensibilidade permanece 50% mas a especificidade é aumentada para 90%. O valor preditivo é agora de 100/180 ou 56%. Então, o aumento na especificidade resultou em aumento muito maior no valor preditivo do que o mesmo aumento na sensibilidade.

Por que a especificidade tem maior efeito que a sensibilidade no valor preditivo? A resposta se torna clara pelo exame desses números. Como estamos lidando com doenças infreqüentes, a maioria da população permanece no lado direito da linha vertical. Conseqüentemente, qualquer mudança à direita da vertical afeta um número maior de pessoas do que afetaria uma mudança comparável à esquerda. Portanto, uma mudança na especificidade possui maior efeito no valor preditivo do que a mudança na sensibilidade. Se lidássemos com uma doença de alta prevalência, a situação seria diferente.

O efeito das mudanças na especificidade no valor preditivo também é visto na Tabela 4–6 de forma similar à usada na Tabela 4–4.

Como foi visto neste exemplo, mesmo com 100% de sensibilidade, uma mudança na especificidade de 70% a 95% teve um efeito dramático no valor preditivo positivo.

CONFIABILIDADE (REPRODUTIBILIDADE) DOS TESTES

Vamos considerar outro aspecto da avaliação diagnóstica e dos testes de rastreamento – a questão é: se um teste é confiável ou deve ser repetido. Podem os resultados obtidos ser reproduzidos se o teste for repetido? Claramente, em vez de sensibilidade ou especificidade de um teste, se o resultado deste não puder ser reproduzido, seu valor e utilidade são mínimos. O restante deste capítulo concentra-se na confiabilidade ou reprodutibilidade dos testes diagnósticos e de rastreamento. Os fatores que contribuem para a variação entre os resultados dos testes são discutidos primeiramente: variação intra-sujeito (variação nos elementos individuais) e variação interobservador (variação entre aqueles que estão lendo os resultados do teste).

Variação Intra-Sujeito

Os valores obtidos na medição de muitas características humanas geralmente variam com o tempo, mesmo durante um período curto. A Tabela 4–7 mostra alterações nas leituras da pressão sangüínea por um período de 24 horas em três indivíduos. A variabilidade através do tempo é considerável. Isso, bem como as condições sob as quais certos testes são conduzidos (p. ex., pós-prandial ou pós-exercício, em casa ou no consultório médico), claramente podem levar a diferentes resultados no mesmo indivíduo. Assim, avaliando qualquer resultado, é importante considerar as condições em que o teste foi realizado, incluindo a hora do dia.

TABELA 4–7. *Exemplos Mostrando a Variação na Pressão Sangüínea Monitorizada por um Período de 24 Horas*

Pressão Sangüínea (mm Hg)	Sexo Feminino Idade 27 Anos	Sexo Feminino Idade 62 Anos	Sexo Masculino Idade 33 Anos
Basal	110/70	132/82	152/109
Menor hora	86/47	102/61	123/78
Maior hora	126/79	172/94	153/107
Casual	108/64	155/93	157/109

De Richardson DW, Honour AJ, Fenton GW, *et al.*: Variação na pressão arterial durante o dia e a noite. Clin Sci 26:445, 1964.

TABELA 4–8. *Variação Instrumental ou do Observador: Concordância em Porcentagem*

	Leitura nº 1			
Leitura nº 2	**Anormal**	**Suspeito**	**Duvidoso**	**Normal**
Anormal	[A]	B	C	D
Suspeito	E	[F]	G	H
Duvidoso	I	J	[K]	L
Normal	M	N	O	[P]

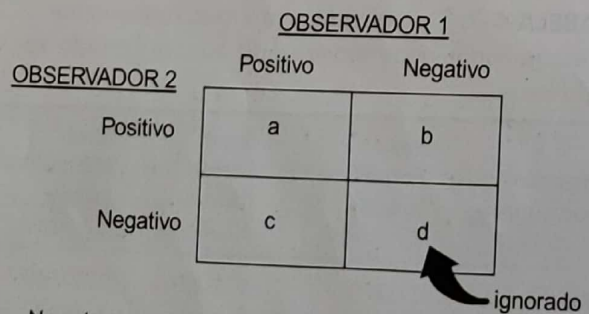
$$\text{Porcentagem de concordância} = \frac{A + F + K + P}{\text{Total de Leituras}} \times 100$$

Variação Interobservador

Outra consideração importante é a variação entre observadores. Dois examinadores geralmente não produzem o mesmo resultado. A extensão na qual os observadores concordam ou não é um importante assunto, se consideramos o exame físico, os testes laboratoriais, ou outros meios de avaliar as características humanas. Desse modo, necessitamos de ser capazes de expressar a extensão da concordância em termos quantitativos.

Concordância Total em Porcentagem

A Tabela 4–8 apresenta um esquema para exame da variação entre observadores. Dois observadores foram instruídos para categorizar cada resultado do teste em uma das quatro categorias: anormal, suspeita, duvidosa ou normal. Este diagrama pode referir-se, por exemplo, a leituras realizadas por dois radiologistas. Neste diagrama, as leituras do observador 1 são cruzadas com as realizadas pelo observador 2. O número de leituras em cada célula é descrito por uma letra do alfabeto. Portanto, os raios X *A* foram lidos como anormais pelos dois radiologistas. Os raios X *C* foram lidos como anormais pelo radiologista 2 e duvidoso pelo radiologista 1. Os raios X *M* foram lidos como anormais pelo radiologista 1 e anormais pelo radiologista 2.



Nas observações em pares, nas quais pelo menos uma das observações de cada par foi positiva, a porcentagem de concordância = $\frac{a}{a+b+c} \times 100$

FIGURA 4–12. Porcentagem de concordância ao se examinar as observações em pares entre o observador 1 e o observador 2.

Como visto na Tabela 4–8, para calcular a porcentagem de concordância total, adicionamos todas as células em que as leituras de ambos os radiologistas concordam *(A + F + K + P)*, dividimos o resultado pelo número total de raios X lidos e multiplicamos o resultado por 100 para transformar em porcentagem.

Em geral, a maioria das pessoas testadas possui resultados negativos. Há uma provável concordância entre os dois observadores em relação a estes sujeitos negativos ou normais. Além disso, quando a porcentagem de concordância é calculada para todos os estudos, seu valor pode ser alto somente porque há grande número de achados negativos em que os observadores concordam. O alto valor pode ocultar incompatibilidade significativa entre os observadores em relação à identificação dos sujeitos como positivos.

Uma abordagem desse problema, observada na Figura 4–12, é desconsiderar os sujeitos rotulados como negativos por ambos os observadores (célula *d*) e calcular a porcentagem de concordância, usando como denominador somente os sujeitos rotulados como normais pelo menos por um observador (células *a*, *b*, ou *c*).

Então, nas *observações em pares* em que pelo menos um dos achados de cada par for positivo, aplica-se a equação a seguir:

$$\text{Porcentagem de concordância} = \frac{a}{a+b+c} \times 100$$

Estatística Kappa

A porcentagem de concordância é também significativamente afetada pelo fato de que, se dois observadores usarem critérios completamente diferentes para denominar os sujeitos como positivos ou negativos, podemos esperar uma concordância sobre alguns elementos somente como uma função de acaso.

Isso pode ser mostrado intuitivamente no seguinte exemplo: você é um diretor de um departamento de radiologia que está lotado e um grande número de raios X deve ser avaliado. Para resolver seu problema, você vai até a rua e pergunta para alguns moradores da vizinhança, que não possuem nenhum conhecimento de biologia ou medicina, para ler os raios X como positivos ou negativos. A primeira pessoa irá até a pilha de raios X e lerá os exames como positivos, negativos, negativos, positivos etc. A segunda pessoa fará o mesmo, da mesma maneira. Visto que essas pessoas não possuem conhecimento, critérios ou padrões para lerem os raios X, uma de suas leituras em um específico raios X estará certa? A resposta é claro que sim; elas irão concordar algumas vezes, mas *puramente pelo acaso*.

Se nós quisermos saber como os dois observadores leram seus exames, poderíamos perguntar: em que extensão suas leituras concordam *além daquilo que esperaríamos somente pelo acaso*? Ou, colocado de outra forma: em que extensão a concordância entre os dois observadores excede o nível de concordância que poderia ser resultado apenas do acaso?

Uma abordagem na resposta dessa questão é calcular a *estatística kappa*, proposta por Cohen em 1960.[2] A estatística kappa pode ser definida pela equação (4.1), mostrada no final da página.

O que o numerador de kappa representa? Queremos saber o quanto a concordância entre as leituras dos observadores é melhor do que aquela que se poderia esperar simplesmente pelo acaso, ou a *porcentagem de concordância observada* menos a *porcentagem de concordância esperada pelo acaso*. Vamos observar o denominador. O 100% no denominador representa concordância total – os dois observadores concordam completamente. O máximo que os observadores poderiam melhorar seus resultados sobre os resultados esperados somente pelo acaso é a diferença entre a *concordância total* e a *porcentagem de concordância esperada somente pelo acaso*, como indicado pelo denominador. Então, o kappa quantifica a extensão em que a concordância observada excede a que seria esperada pelo acaso e expressa isto como a proporção de melhoria máxima que poderia ocorrer além da concordância esperada pelo simples acaso, que os observadores obtiveram.

Para calcular o kappa, devemos primeiro calcular a quantidade de concordância que deveria ser esperada simplesmente na base do acaso. Considere dados relatados na classificação histológica de câncer de pulmão que focaliza a capacidade de reprodução, classificando em subtipos de grandes células de carcinoma de pulmão.[3]

A Figura 4–13 mostra dados comparando os achados de dois patologistas classificando esses casos em subtipos.

A primeira questão é: qual a concordância observada entre os dois patologistas? A Figura 4–14 mostra as leituras do patologista A para todo o grupo de que estão na parte inferior da mesa e as do patologista B na porção direita da mesa. Então, o patologista A identificou 45 ou 60% de todas as 75 biópsias como grau II e o patologista B identificou 44 ou 58,6% de todas as biópsias como grau II. Como discutido no início do capítulo, a porcentagem de concordância é determinada pela seguinte equação:

$$\text{Porcentagem de concordância} = \frac{41 + 27}{75} \times 100 = 90{,}7\%$$

Isto é, os patologistas concordaram em 90,7% das leituras.

A próxima questão é: se os dois patologistas utilizaram diferentes critérios, quanto de concordância seria obtida somente na base do acaso? O patologista A leu 60% de todas as 75 biópsias (45 biópsias) como sendo grau II. Se as suas leituras usassem os critérios independente dos usados pelo patologista B *(i. e., se o patologista A fosse ler 60% de qualquer grupo de biópsias como sendo de grau II), poderíamos esperar que o patologista A leria como grau II ambos os 60% das biópsias que o patologista B denominou como grau II e 60% das biópsias* que o patologista B determinou como grau III. Então esperaríamos que 60% (26,4) das 44 biópsias determinadas como grau II pelo patologista B poderiam ser consideradas de grau II pelo patologista A, e que 60% (18,6) das 31 biópsias consideradas como grau III pelo patologista B poderiam também ser consideradas como sendo de grau II pelo patologista A (Fig. 4–15).

(Equação 4.1)

$$\text{Kappa} = \frac{\text{(Porcentagem da concordância observada)} - \text{(Porcentagem da concordância esperada somente pelo acaso)}}{100\% - \text{(Porcentagem da concordância esperada somente pelo acaso)}}$$

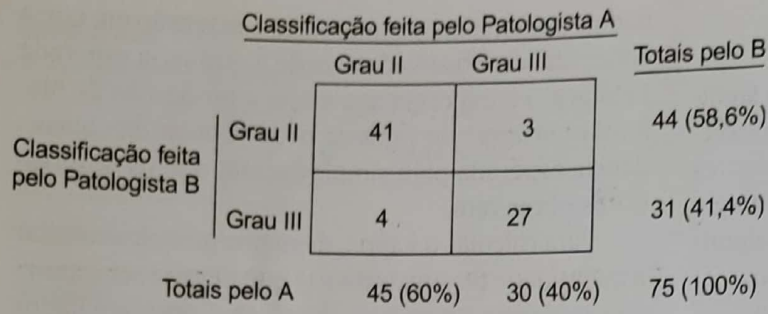


FIGURA 4–13. Classificação histológica pelo subtipo de 75 biópsias de carcinoma de grandes células, por dois patologistas (A e B). (Dados de Ghandur-Mnaymneh L, Raub WA, Sridhar KS, *et al.*: A exatidão da classificação do carcinoma de pulmão e sua capacidade de reprodução: Um estudo de 75 casos arquivados de carcinoma adenoescamoso. Cancer Invest 11:641, 1993.)

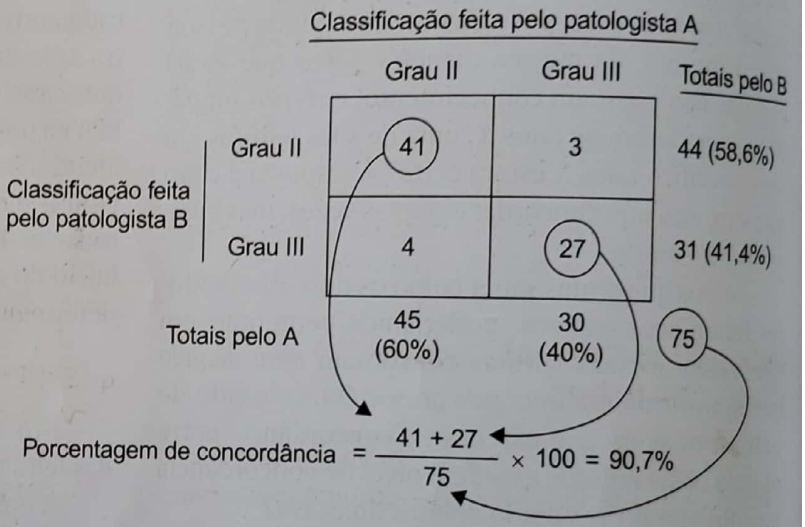


FIGURA 4–14. Porcentagem de concordância pelo patologista A e patologista B. (Dados de Ghandur-Mnaymneh L, Raub WA, Sridar KS, *et al.*: A exatidão da classificação histológica no carcinoma de pulmão e sua reprodutibilidade: Um estudo de 75 casos arquivados de carcinoma adenoescamoso. Cancer Invest 11:641, 1993.)

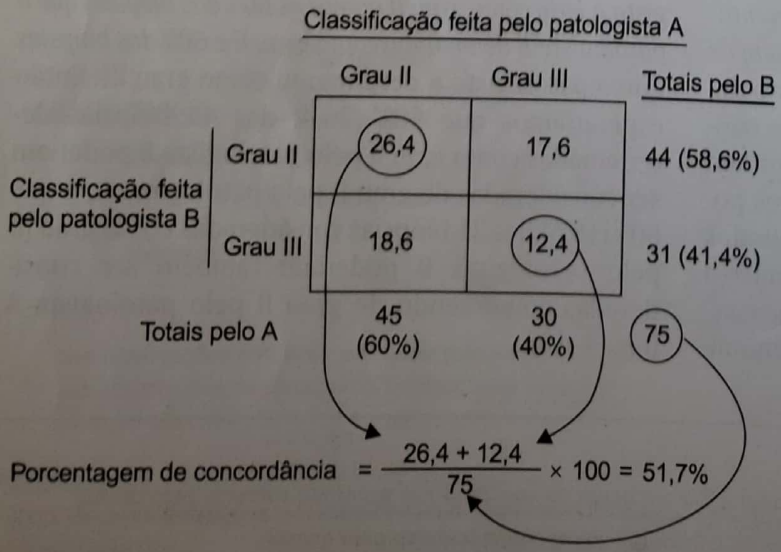


FIGURA 4–15. Porcentagem de concordância pelo patologista A e patologista B esperada somente pelo acaso. (Dados de Ghandur-Mnaymneh L, Raub WA, Sridar KS, *et al.*: A exatidão da classificação histológica no carcinoma de pulmão e sua reprodutibilidade: Um estudo de 75 casos arquivados de carcinoma adenoescamoso. Cancer Invest 11:641, 1993.)

(Equação 4.2)

$$\text{Kappa} = \frac{\text{(Porcentagem da concordância observada)} - \text{(Porcentagem da concordância esperada somente pelo acaso)}}{100\% - \text{(Porcentagem da concordância esperada somente pelo acaso)}}$$

$$= \frac{90,7\% - 51,7\%}{100\% - 51,7\%} = \frac{39\%}{48,3\%} = 0,81$$

Então, a concordância esperada somente pelo acaso seria $\frac{26,4+12,4}{75} = \frac{38,8}{75} = 51,7\%$ de todas as biópsias lidas. O kappa pode então ser calculado usando-se a mesma fórmula (equação 4.2) mostrada no topo da página.

Landis e Koch[4] sugeriram que o kappa maior que 0,75 representa uma excelente concordância além do acaso, que o kappa abaixo de 0,40 representa uma concordância baixa, e que o kappa entre os valores de 0,40 e 0,75 representa uma concordância intermediária razoável. O teste para a significância estatística do kappa é descrito por Fleiss.[5] Discussões consideráveis ocorreram em relação ao uso apropriado do kappa, matéria estudada por MacLure e Willett.[6]

RELAÇÃO ENTRE VALIDADE E CONFIABILIDADE

Para concluir este capítulo, vamos comparar a validade e a confiabilidade, usando uma representação gráfica.

A linha horizontal da Figura 4–16 é uma escala de valores para uma determinada variável, como o nível de glicose do sangue, com os seus valores verdadeiros indicados. Os resultados dos testes obtidos são mostrados pela curva. A curva é estreita, indicando que os resultados são bastante confiáveis (repetitivos); infelizmente, entretanto, eles agrupam-se longe do valor verdadeiro, portanto não são válidos. A Figura 4–17 mostra uma curva ampla e, assim, possui baixa confiabilidade. Mas os valores obtidos agrupam-se ao redor do valor real e são, portanto, válidos. Claramente, o que gostaríamos de obter são resultados tanto válidos como confiáveis (Fig. 4–18).

É importante salientar que, na Figura 4–17, na qual a distribuição dos resultados dos testes é uma curva ampla centrada no valor verdadeiro, e digamos que os resultados sejam válidos, eles somente o serão para um grupo *(i. e.*, eles tendem a agrupar-se ao redor do valor verdadeiro). Entretanto, o que pode ser válido para um grupo ou para uma população pode não ser para um indivíduo na clínica. Quando a confiabilidade ou repetitividade do teste é pobre, a validade deste para um certo indivíduo pode também ser pobre. A distinção entre a validade do grupo e a validade individual é, portanto, um fator importante para

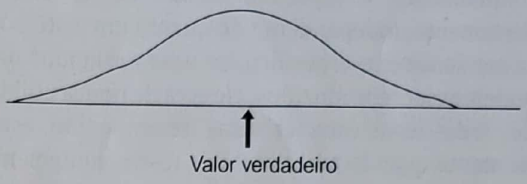


FIGURA 4–17. Gráfico dos resultados hipotéticos de testes que são válidos, mas não confiáveis

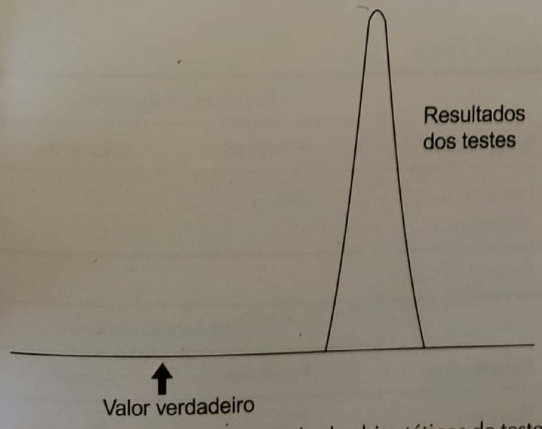


FIGURA 4–16. Gráfico dos resultados hipotéticos de testes que são confiáveis, mas inválidos.

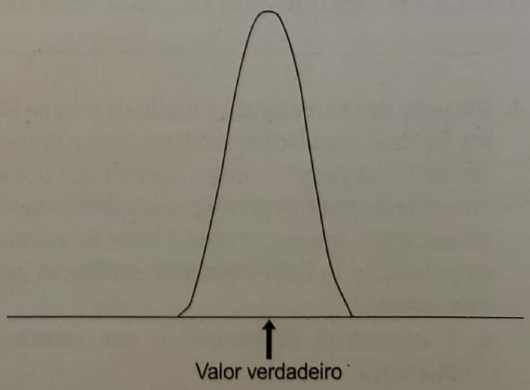


FIGURA 4–18. Gráfico dos resultados hipotéticos de testes que são confiáveis e válidos.

lembrar quando se avalia a qualidade dos testes diagnósticos e de rastreamento.

CONCLUSÃO

Este capítulo discutiu a validade dos testes diagnósticos e de rastreamento, medidos pela sua sensibilidade e especificidade, seu valor preditivo e a confiabilidade e reprodutibilidade desses testes. Certamente, independente do quanto um teste possa ser sensível e específico, se seus resultados não puderem ser reproduzidos, ele será de pouca utilidade. Todas essas características devem, então, estar na mente quando avaliamos tais testes, juntamente com os objetivos para os quais o teste vai ser realizado.

REFERÊNCIAS

1. De Sheffield LJ, Sackett DL, Goldsmith CH, *et al.*: Uma abordagem clínica para o uso dos valores preditivos no diagnóstico pré-natal dos defeitos do tubo neural. Am J Obstet Gynecol 1465:319, 1983.)
2. Cohen J: Um coeficiente de concordância para escalas nominais. Educ Psychol Meas 20:37, 1960.
3. Ghandur-Mnaymneh L, Raub WA, Sridhar KS, *et al.*: A exatidão da classificação histológica de carcinoma de pulmão e sua capacidade de reprodução: um estudo de 75 casos arquivados de carcinoma adenoescamoso. Cancer Invest 11:641, 1993.
4. Landis JR, Koch GG: A medição da concordância do observador para dados por categoria. Biometrics 33:159, 1977.
5. Fleiss JL: Métodos Estatísticos para Índices e Proporções, ed 2. New York, John Wiley & Sons, 1981.
6. MacLure M, Willet WC: Interpretação equivocada e uso indevido da estatística kappa. Am J Epidemiol 126:161, 1987.

Questões de Revisão para o Capítulo 4

As perguntas 1, 2 e 3 são baseadas nas informações fornecidas abaixo:

Um exame físico foi utilizado para rastrear 2.500 mulheres quanto a câncer de mama comprovado por biópsia e em 5.000 mulheres (controle), classificadas por idade e raça. Os resultados de um exame físico foram positivos (*i. e.*, a massa foi palpada) em 1.800 casos e em 800 casos-controle, em que não houve evidência de câncer na biópsia.

1. A sensibilidade do exame físico foi:

   _____

2. A especificidade do exame físico foi:

   _____

3. O valor preditivo positivo do exame físico foi:

   _____

4. Um teste de rastreamento é usado da mesma forma em duas populações similares, mas a proporção de falsos positivos entre aqueles que possuem um teste positivo para a população A é menor do que entre aqueles em que o teste foi positivo na população B. Qual a possível explicação para este achado?
   a. É impossível determinar o que causou a diferença
   b. A especificidade do teste é menor na população A
   c. A prevalência da doença é menor na população A
   d. A prevalência da doença é maior na população A
   e. A especificidade do teste é maior na população A

A pergunta 5 é baseada nas seguintes informações:

Um exame físico e um teste audiométrico foram realizados em 500 pessoas com suspeita de problemas auditivos, em que 300 foram detectados. O resultado dos exames são:

Exame Físico

	Problemas Auditivos	
Resultados	**Presente**	**Ausente**
Positivo	240	40
Negativo	60	160

Teste Audiométrico

	Problemas Auditivos	
Resultados	**Presente**	**Ausente**
Positivo	270	60
Negativo	30	140

5. Comparado com o exame físico, o teste audiométrico é:
   a. Igualmente sensível e específico
   b. Menos sensível e menos específico
   c. Menos sensível e mais específico
   d. Mais sensível e menos específico
   e. Mais sensível e mais específico

A pergunta 6 é baseada nas seguintes informações:

Dois pediatras querem investigar um novo teste laboratorial que identifica infecções estreptocócicas. Dr. Kidd usa um teste de cultura padrão, que tem uma sensibilidade de 90% e uma especificidade de 96%. Dr. Childs usa um novo teste, que é 96% sensível e 96% específico.

6. Se 200 pacientes forem submetidos à cultura com ambos os testes, qual das respostas é correta?
   a. Dr. Kidd identificará corretamente mais pessoas *com* infecções estreptocócicas do que o Dr. Childs
   b. Dr. Kidd identificará corretamente menos pessoas *com* infecções estreptocócicas do que o Dr. Childs
   c. Dr. Kidd identificará corretamente mais pessoas *sem* infecções estreptocócicas do que o Dr. Childs
   d. A prevalência de uma infecção estreptocócica *é* necessária para determinar qual pediatra irá identificar corretamente maior número de pessoas com a doença

As perguntas 7 e 8 são baseadas nas seguintes informações:

Um estudo de rastreamento para o câncer de cólon está sendo conduzido em Nottingham, Inglaterra. Indivíduos de 50 a 75 anos serão rastreados com o teste Hemoccult. Nesse teste, uma amostra de fezes é testada quanto à presença de sangue.

7. O teste Hemoccult tem uma sensibilidade de 70% e uma especificidade de 75%. Se Nottingham possui uma prevalência de 12/1.000 para câncer de cólon, qual o *valor preditivo positivo* do teste?____

8. Se o teste Hemoccult for negativo, nenhum outro teste será realizado. Se o teste Hemoccult for positivo, o indivíduo deverá realizar novo exame de fezes testado com Hemoccult II. Se esta segunda amostra também for positiva para sangue, o indivíduo será mais bem investigado. O efeito na sensibilidade e na especificidade líquidas deste método de rastreamento é:
   a. Sensibilidade líquida e especificidade líquida estão aumentadas
   b. a sensibilidade líquida está diminuída e a especificidade líquida está aumentada
   c. A sensibilidade líquida permanece a mesma e a especificidade líquida está aumentada
   d. A sensibilidade líquida está aumentada e a especificidade líquida está diminuída
   e. O efeito da sensibilidade líquida e a especificidade líquida não pode ser determinado pelos dados

As perguntas 9 a 12 são baseadas nas informações fornecidas abaixo:

Foi solicitado a dois médicos para classificarem 100 raios X de tórax como "anormal" ou "normal" à sua escolha. A comparação de suas classificações é mostrada na tabela:

Classificação de Raios X de Tórax pelo Médico 1, Comparado com o Médico 2

	Médico 2		
Médico 1	**Anormal**	**Normal**	**Total**
Anormal	40	20	
Normal	10	30	40
Total	50	50	100

9. A porcentagem de concordância geral e simples entre os dois médicos foi de:_____

10. A porcentagem geral de concordância entre os dois médicos, retirando os raios X classificados como normais por ambos os médicos, é:_____

11. O valor kappa é:_____

12. Este kappa representa que tipo de concordância?
    a. Excelente
    b. Intermediária a boa
    c. Pobre