Epidemiologia Aplicada à Clínica

Capítulo de Diagnóstico

Antonio Alberto Lopes

Marcelo Barreto Lopes

Autores

Antonio Alberto Lopes, MD, PhD, MPH, MSc

Médico graduado pela Faculdade de Medicina da Bahia, Universidade Federal da Bahia (UFBA). Professor Titular, Livre Docente, Faculdade de Medicina da Bahia, Universidade Federal da Bahia (UFBA). Professor de Clínica Médica, Medicina Baseada em Evidência e Epidemiologia Aplicada à Clínica. Professor Permanente do Programa de Pós-Graduação em Medicina e Saúde, UFBA. Chefe do Núcleo de Epidemiologia Clínica e Medicina Baseada em Evidências do Hospital Universitário Professor Edgard Santos, UFBA. PhD em Ciência Epidemiológica e MPH pela School of Public Health, Universidade de Michigan, Ann Arbor, MI, Estados Unidos. Mestrado em Medicina Interna, UFBA. Pesquisador do CNPq. Membro Titular da Academia de Medicina da Bahia.

Marcelo Barreto Lopes, MD, PhD, MPH

Médico graduado pela Faculdade de Medicina da Bahia, Universidade Federal da Bahia (UFBA). Especialista em Nefrologia pela Sociedade Brasileira de Nefrologia. Doutor em Medicina e Saúde pelo Programa de Pós-Graduação em Medicina e Saúde, UFBA. Epidemiologista com MPH pela School of Public Health, Universidade de Michigan, Ann Arbor, MI, Estados Unidos. Pesquisador na Instituição de Pesquisa Arbor Research Collaborative for Health, Ann Arbor, MI, Estados Unidos. Professor de Cursos de Medicina Baseada em Evidência.

Testes para o Diagnóstico Diferencial e Detecção Precoce de Doenças

CAPÍTULO 6

Na pesquisa clínico-epidemiológica, o termo teste diagnóstico se refere a sintomas, sinais do exame físico, dados laboratoriais, dados de exame de imagem e outros tipos de dados usados no diagnóstico diferencial e no rastreamento (*screening*) para detecção precoce de doenças. Os termos teste, dado e achado são usados indistintamente neste capítulo para se referir a um dado do exame clínico ou um exame complementar usado para o diagnóstico.

Existem importantes aspectos sobre testes diagnósticos que devem ser considerados: 1) a acurácia ao ser comparado com um teste padrão-ouro do diagnóstico de uma condição; 2) a capacidade de predizer, ou seja, aumentar ou reduzir a probabilidade de um diagnóstico; 3) a incerteza ou o limite de certeza da estimativa representado pelo intervalo de confiança; e 4) a reprodutibilidade.

O padrão-ouro

No contexto do diagnóstico clínico, o padrão-ouro (referência padrão, critério padrão; *gold standard*, em inglês) é um método, procedimento, critério ou teste considerado o melhor para identificar quem tem ou não um determinado problema de saúde ou doença. O padrão-ouro serve de base para estimar a acurácia de um teste diagnóstico, ou seja, o grau em que o teste representa a presença ou ausência do evento. Quando o padrão-ouro já é um teste simples e de baixo custo, talvez não seja necessário avaliar outros testes. Contudo, o padrão-ouro é, frequentemente, um teste de alto custo, de elevado risco ou que necessita de tecnologia

que não é largamente disponível e profissionais adequadamente treinados. Em algumas circunstâncias, o padrão-ouro requer um tempo longo para definir quem tem ou não uma doença.

As células a, b, c, d da Tabela 1 mostram os possíveis resultados de um teste diagnóstico classificado em dois níveis (positivo e negativo), quando comparado com o diagnóstico da doença pelo padrão-ouro. Quando o teste é apresentado em dois níveis, ao ser comparado com o padrão-ouro, dois resultados do teste são considerados corretos (verdadeiros) e dois, incorretos (falsos). O teste é considerado verdadeiro quando for positivo na presença da doença (célula a) ou quando for negativo na ausência de doença (célula d). Contrariamente, o teste é considerado falso quando for negativo na presença da doença (célula c) ou positivo na ausência da doença (célula b).

Tabela 1. Relação entre o teste diagnóstico e a doença pelo padrão-ouro.

		Doença pelo padrão-ouro		
		Presente	Ausente	Totais
Teste diagnóstico	Positivo	a Verdadeiro	b Falso	a + b
	Negativo	c Falso	d Verdadeiro	c + d
	Totais	a + c	b + d	a + b + c + d

Fonte: Autores.

Prevalência e chances (*odds*) da doença

Prevalência total (prevalência pré-teste, prevalência verdadeira)

A prevalência total é representada pela proporção de casos da doença diagnosticados pelo padrão-ouro no grupo total. A prevalência total tem sido também denominada prevalência pré-teste e prevalência verdadeira, para diferenciar da prevalência aparente.

$$\text{Prevalência total} = \frac{(a + \mathbf{c})}{(a + \mathbf{b} + \mathbf{c} + d)}$$

Para que o estudo propicie estimativa válida da prevalência, a seleção da amostra deve ser baseada em um espectro de pacientes com um problema clínico similar ao observado na prática; por exemplo, uma amostra de pacientes adultos com dor abdominal aguda atendidos em serviços de emergências com o objetivo de investigar o valor de um teste no diagnóstico de pancreatite aguda. Se o estudo inclui todos os pacientes com dor abdominal aguda atendidos em alguns serviços de emergência em um determinado período, e 42 de 600 pacientes dessa amostra forem diagnosticados como portadores de pancreatite aguda pelo padrão-ouro, então a prevalência será de 7%. O estudo não propicia uma estimativa válida de prevalência se a relação doente/não doente difere do que ocorre na população clínica. No exemplo de pacientes com dor abdominal, o estudo não irá propiciar uma estimativa válida da prevalência se forem selecionados 42 pacientes com diagnóstico confirmado de pancreatite aguda e como controle 84 pacientes sem pancreatite aguda, considerando que essa não é a relação entre os diagnósticos em pacientes com dor abdominal aguda atendidos em serviços de emergência.

Prevalência aparente

Como o termo já diz, não é a verdadeira prevalência da doença. Ela é representada pelo percentual de indivíduos com teste positivo (verdadeiro-positivo e falso-positivo) em todos os indivíduos estudados.

$$\text{Prevalência total} = \frac{(a + \mathbf{b})}{(a + \mathbf{b} + \mathbf{c} + d)}$$

A prevalência aparente é o que é visto na prática clínica quando a ocorrência da doença é baseada no resultado do teste diagnóstico. Usando a prevalência aparente em conjunto com a sensibilidade e a especificidade do teste, é possível estimar a prevalência verdadeira usando a equação a seguir em que os valores da prevalência aparente, especificidade e sensibilidade são usados como proporção em lugar de percentual. Deve-se multiplicar o resultado por 100 para obter a prevalência em percentual.

$$\text{Prevalência verdadeira} = \frac{(\text{prevalência aparente} + \textbf{especificidade} - 1)}{(\textbf{especificidade} + \text{sensibilidade} - 1)}$$

Chances (*odds*) da doença

Chances (sim, no plural) é o termo em português para *odds* e são representadas pela razão entre uma proporção e seu complemento. No contexto do diagnóstico, as chances da doença são determinadas dividindo a prevalência da doença pelo seu complemento. Exemplo: se a prevalência for 25%, as chances serão 25% para 75% (25/75). As chances podem ser também representadas pela razão entre o número de um atributo estar presente e o número do mesmo atributo estar ausente. Como mostrado abaixo, as chances (*odds*) da doença são representadas pela divisão do número de doentes pelo número de não doentes, de acordo com o padrão-ouro.

$$\text{Chances da doença} = \frac{(a + \textbf{c})}{(\textbf{b} + d)}$$

Um resultado expresso como chances pode ser transformado em proporção (ex., prevalência) usando a fórmula abaixo. Deve-se multiplicar por 100 para apresentar o resultado em percentual.

$$\text{Proporção (exemplo: Prevalência)} = \frac{\text{Chances}}{1 + \text{Chances}}$$

Medidas de testes diagnósticos classificadas em dois níveis

A estrutura da Tabela 1 serve de base para os cálculos de medidas de testes diagnósticos classificados em dois níveis: sensibilidade, especificidade, *likelihood ratio* (LR) positiva, LR negativa e valores preditivos.

Sensibilidade

O cálculo da sensibilidade é feito no eixo vertical da tabela considerando apenas a coluna da doença presente, de acordo com o padrão-ouro.

A sensibilidade representa o percentual de testes positivos verdadeiros (célula a) entre os doentes pelo padrão-ouro (soma das células a e c).

$$\text{Sensibilidade} = \frac{a}{a + \mathbf{c}}$$

Especificidade

O cálculo da especificidade é feito também no eixo vertical da tabela. A especificidade representa o percentual de testes negativos verdadeiros (célula d) entre os indivíduos com doença ausente pelo padrão-ouro (soma das células b e d).

$$\text{Especificidade} = \frac{d}{\mathbf{b} + d}$$

Likelihood ratio (razão de verossimilhança)

A *likelihood ratio* tem sido traduzida para o português usando diferentes termos, como razão de verossimilhança e razão de probabilidades. Neste capítulo são usados o termo original em inglês (*likelihood ratio*) e a sigla LR por ser o que é usado em artigos publicados e para facilitar a comunicação com um único termo e uma sigla fácil de memorizar. Como é mostrado no capítulo sobre medicina baseada em evidências, a LR é uma medida prática para transformar a probabilidade pré-teste de um diagnóstico em probabilidade pós-teste.

Quando o teste é classificado em dois níveis, as LRs são denominadas positiva (LR+) e negativa (LR-). Similarmente à sensibilidade e à especificidade, as LRs são calculadas usando os números no eixo vertical da tabela. Diferentemente da sensibilidade e especificidade, a presença e ausência da doença são consideradas conjuntamente tanto no cálculo da LR+ quanto da LR-. A LR para um determinado nível de um teste diagnóstico demonstra quantas vezes maior ou menor é a proporção de testes naquele nível em pacientes com a doença de interesse em relação a pacientes sem a doença, no mesmo nível do teste.

Likelihood ratio positiva

A LR+ é calculada pela razão entre a proporção de testes positivos nos indivíduos com a doença e a proporção de testes positivos nos indivíduos sem a doença. Esse cálculo é equivalente à divisão da sensibilidade pelo complemento da especificidade.

$$LR+ = \frac{\left(\frac{a}{(a + \mathbf{c})}\right)}{\left(\frac{\mathbf{b}}{(\mathbf{b} + d)}\right)} \qquad LR+ = \frac{\text{sensibilidade}}{\text{complemento da especificidade}}$$

Likelihood ratio negativa

A LR- é representada pela razão entre a proporção de testes negativos entre indivíduos com a doença e a proporção de teste negativos em indivíduos sem a doença. Esse cálculo é equivalente à divisão do complemento da sensibilidade pela especificidade.

$$LR- = \frac{\left(\frac{\mathbf{c}}{(a + \mathbf{c})}\right)}{\left(\frac{\mathbf{d}}{(\mathbf{b} + d)}\right)} \qquad LR- = \frac{\text{complemento da sensibilidade}}{\text{especificidade}}$$

Valores preditivos

O valor preditivo positivo (VPP) e o valor preditivo negativo (VPN) medem a proporção de valores verdadeiros em indivíduos com teste positivo e negativo, respectivamente. Por serem calculados no eixo horizontal da tabela, ou seja, nas linhas em vez de nas colunas, os valores preditivos variam, dependendo da prevalência total da doença. As fórmulas do VPP e VPN são mostradas em seguida.

$$\text{Valor preditivo positivo (VPP)} = \frac{a}{a + \mathbf{b}} \qquad \text{Valor preditivo negativo (VPN)} = \frac{d}{\mathbf{c} + d}$$

Estimativas pontuais e intervalos de confiança de medidas de teste diagnóstico

As estimativas pontuais de medidas de teste diagnóstico são os valores calculados ao transferir os números observados e os totais para as fórmulas apresentadas anteriormente. O intervalo de confiança (IC) 95% é o indicador da precisão de uma estimativa mais frequentemente usado. O IC 95% informa com 95% de certeza os limites de valores em que se encontra o verdadeiro valor da medida.

Ao avaliar os resultados de medidas de teste diagnóstico, é interessante observar os limites do IC em conjunto com a estimativa pontual. A medida é considerada de maior contribuição no processo diagnóstico quanto mais relevante for o limite do IC de menor valor para o diagnóstico. Para a sensibilidade, a especificidade e os valores preditivos, o limite inferior do IC é o de menor valor diagnóstico. Similarmente, o limite inferior da LR+ é o de menor valor para confirmar a presença da doença. Contrariamente, para a LR- é o limite superior que propicia menor contribuição para o diagnóstico, ou seja, para afastar a doença.

Os dados nas tabelas 2 e 3 são usados como exemplos de estimativas pontuais e IC 95% de medidas de testes diagnósticos. Os dados são de um estudo que investigou o valor da relação proteína pleura/soro (RPPS) no diagnóstico diferencial do derrame pleural entre exsudato e transudato em uma amostra de 150 pacientes.[1] No exemplo, exsudato é o tipo de derrame pleural de maior interesse, pois quando presente será necessário o prosseguimento da avaliação para o diagnóstico da causa, como câncer ou infecção, e transudato é a comparação. A RPPS foi classificada como positiva se superior a 0,5 e negativa se ≤ 0,5. Para definir o padrão-ouro, foram definidos critérios antes do início do estudo para diagnosticar se a condição era causadora de exsudato ou transudato.

Tabela 2. Razão proteína/soro como indicadora do tipo de derrame pleural.

		Tipo de derrame pleural		
		Exsudato	Transudato	Totais
Razão proteína pleura/soro	> 0,5	**93**	**1**	94
	≤ 0,5	**10**	**46**	56
	Totais	103	47	150

Fonte: Light.[1]

Como mostrado na Tabela 3, a estimativa pontual da sensibilidade foi 90,3% (93/103*100%). Se tiver que referir apenas um valor para a sensibilidade da RPPS, então 90,3% é o valor que deve ser referido. Contudo, ao considerar a precisão da estimativa expressa no IC 95%, a conclusão com 95% de certeza é que o verdadeiro valor da sensibilidade é algum valor entre 82,9% (o limite inferior) e 95,3% (o limite superior). Quanto à especificidade, os resultados mostram que a estimativa pontual da RPPS foi de 97,9% (46/47*100%) e com 95% de certeza que a especificidade é algum valor entre 88,7% e próximo a 100%. Tanto o limite inferior do IC 95% da sensibilidade quanto da especificidade foram superiores a 82%, indicando boa contribuição no processo diagnóstico.

Tabela 3. Estimativas pontuais e intervalos de confiança 95%.

Medidas	Estimativas Pontuais	IC* 95%
Sensibilidade	$\frac{93}{103}$ = **90,3%**	82,9% - 95,3%
Especificidade	$\frac{46}{47}$ = **97,9%**	88,7% - 99,95%
Likelihood ratio positiva	$\left(\frac{93}{103}\right) \div \left(\frac{1}{47}\right)$ = **42,4**	6,10 - 295,3
Likelihood ratio negativa	$\left(\frac{10}{103}\right) \div \left(\frac{46}{47}\right)$ = **0,099**	0,06 - 0,18
Prevalência no total	$\frac{103}{150}$ = **68,7%**	60,6% - 80,0%
Valor preditivo positivo	$\frac{93}{94}$ = **98,9%**	94,2% - 99,97%
Valor preditivo negativo	$\frac{10}{56}$ = **82,1%**	69,6% - 91,1%

*IC: intervalo de confiança.

Fonte: Autores.

A LR+ de 42,4 foi determinada pela divisão da proporção do teste positivo no grupo de pacientes com exsudato (sensibilidade), i.e., 0,9029 (93/103) com a proporção do teste positivo no grupo de pacientes sem exsudato (complemento da especificidade), i.e., 0,0213 (1/47). A estimativa pontual de 42,4 da LR+ é um valor expressivo para confirmação

do diagnóstico de exsudato no caso de RPPS > 0,5. Contudo, a precisão da estimativa é pequena conforme mostrado pelo IC 95% muito largo, indo de 6,10 a 295,3. Esse intervalo muito largo do IC 95% é explicado pelo pequeno número de pacientes com transudato (n = 47), particularmente o pequeno número de falso-positivos (apenas 1 com diagnóstico de transudato). Apesar da baixa precisão da estimativa, o limite inferior do IC 95% da LR+ (6,1) é de valor moderado no processo diagnóstico, ou seja, para aumentar a probabilidade de um diagnóstico no caso de teste positivo.[2] Para o cálculo da LR-, as proporções de testes negativos em pacientes com exsudato e transudato (especificidade) foram usadas. A divisão das proporções 0,0971 e 0,9787 resultou numa estimativa pontual da LR- da RPPS de 0,099 — mostrada na Tabela 3. O IC 95% indica com 95% de certeza que a LR- é algum valor entre 0,06 e 0,18. Enquanto o valor inferior do IC 95% da LR- é um valor expressivo para afastar um diagnóstico, o valor superior do IC 95% da LR- é de valor moderado.

São também mostrados na Tabela 3 a prevalência total, o VPP e o VPN, com respectivos intervalos de confiança 95%. Como o VPP e o VPN são dependentes da prevalência da doença no estudo, essas medidas não devem ser vistas como indicadoras do valor do teste diagnósticos.

Balanço entre sensibilidade e especificidade

Para testes representados por uma variável quantitativa, o ponto de corte para definir o que é normal ou anormal exige uma decisão arbitrária. Os dados da Tabela 4 são de uma metanálise que avaliou o valor da ferritina sérica no diagnóstico da anemia por deficiência de ferro.[3]

Tabela 4. Sensibilidade e especificidade usando diferentes pontos de corte da ferritina sérica para o diagnóstico de anemia por deficiência de ferro.

Ponto de Corte da Ferritina (µg/L)	Sensibilidade (%)	Especificidade (%)
< 25	73,1	97,4
< 45	87,4	92,3
< 100	94,1	71,3

Fonte: Adaptado de Guyatt.[3]

Se o ponto de corte para ferritina fosse menor que 25 µg/L, a sensibilidade ficaria em torno de 73% e a especificidade aproximadamente 97%. Por outro lado, se o ponto de corte para nível baixo de anemia ferropriva fosse < 100, o percentual da especificidade seria reduzido de aproximadamente 97% para 71%, enquanto a sensibilidade aumentaria de aproximadamente 73% para 94,1%.

Curva ROC

O balanço entre sensibilidade e especificidade tem sido mostrado com o uso da *receiver operating characteristic curve* (curva ROC). Para demonstrar como é construída a curva ROC, são aqui usados os dados mostrados na Tabela 5: percentuais da sensibilidade, da especificidade e de complementos da especificidade (100 - especificidade) do antígeno prostático específico (PSA) em norte-americanos negros com idade entre 60 a 69 anos; 156 com câncer de próstata e 604 sem evidência de câncer de próstata.[4]

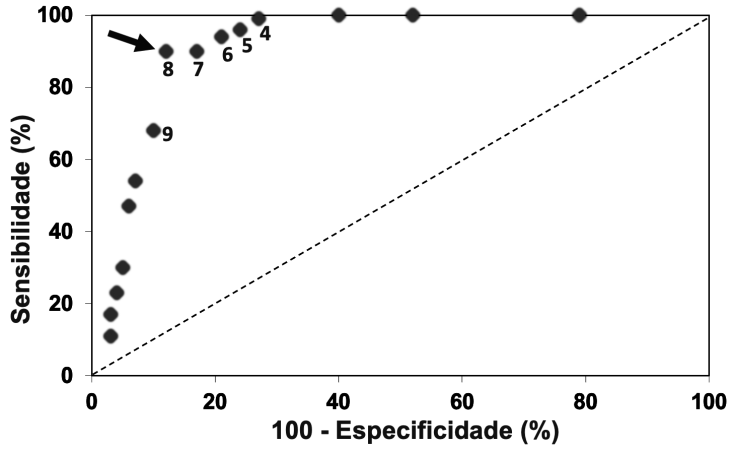
Tabela 5. Percentuais da sensibilidade, da especificidade e do complemento da especificidade (falso-positivos) por nível do PSA sérico.

PSA (ng/mL)	Sensibilidade	Especificidade	100 - especificidade
1	100	21	79
2	100	48	52
3	100	60	40
4	99	73	27
5	96	76	24
6	94	79	21
7	90	83	17
8	90	88	12
9	68	90	10
10	54	93	7
11	47	94	6
12	30	95	5
13	23	96	4
14	17	97	3
15	11	97	3

Fonte: Morgan.[4]

A Figura 1 mostra a curva ROC dos dados da Tabela 5. A curva ROC consiste em um diagrama representando a sensibilidade no eixo Y e o complemento da especificidade (falso-positivo) no eixo X. Curvas de testes com melhor capacidade discriminativa se afastam do centro e ficam mais próximas do ângulo superior esquerdo. Curvas de testes com menor capacidade discriminativa ficam próximas da linha diagonal (linha pontilhada) que vai da parte inferior à esquerda do gráfico até a parte superior do lado direito do gráfico. A área abaixo da curva é usada para indicar o desempenho do teste. Quanto maior a área, melhor é considerado o teste. A área abaixo da curva para esses dados foi de 0,91 (91%). O melhor ponto de corte é considerado o que fica mais próximo do ângulo superior esquerdo do gráfico, que representa a situação em que a soma da sensibilidade e especificidade alcança o maior valor. O nível de PSA que corresponde ao melhor balanço entre sensibilidade e especificidade foi 8 ng/mL. Os números de 4 a 9 identificam valores do PSA em cada ponto da curva. A seta aponta para o nível com melhor balanço entre sensibilidade e especificidade.

Figura 1. Curva ROC de PSA comparando pacientes com e sem câncer de próstata.



Fonte: Adaptado de Morgan.[4]

A curva ROC tem sido usada para comparar desempenho de testes no diagnóstico de uma doença. O teste com maior capacidade de separar os doentes dos não doentes ocupa uma área maior sob a curva.

Uma limitação da curva ROC é que não é possível um ponto de corte que aumente tanto a capacidade do teste diagnóstico de afirmar a presença da doença no caso de teste positivo e de afastar a doença no caso de teste negativo. Se o tamanho amostral permitir, uma alternativa para aumentar o valor diagnóstico de um teste diagnóstico representado por variável quantitativa é apresentar resultados em múltiplos níveis. Desta forma é possível calcular *likelihood ratios* para mais de dois níveis do teste, sem a necessidade de determinar quando o teste deve ser considerado positivo ou negativo.

Teste diagnóstico em múltiplos níveis

A Tabela 6 mostra dados da razão proteína pleura/soro de uma metanálise de dados individuais de 1393 pacientes com diagnóstico etiológico de doença exsudativa e transudativa.[5] Os dados são agrupados em 10 categorias de razão proteína pleura/soro (RPPS), de forma a permitir determinar LR para cada categoria da RPPS.

Tabela 6. *Likelihood ratios* e intervalo de confiança 95% da razão proteína pleura/soro.

Razão proteína pleura/soro	Etiologia do derrame		*Likelihood ratio*	
	Exsudato	Transudato	Estimativa pontual	IC* 95%
> 0,70	475	1	(475/1028)/(1/365) = **169**	23,8; 1195
0,66-0,70	150	1	(150/1028)/(1/365) = **53,3**	7,48; 379
0,61-0,65	117	6	(117/1028)/(6/365) = **6,92**	3,07; 15,6
0,56-0,60	102	12	(102/1028)/(12/365) = **3,02**	1,68; 5,42
0,51-0,55	70	14	(70/1028)/(14/365) = **1,78**	1,01; 3,11
0,46-0,50	47	34	(47/1028)/(34/365) = **0,49**	0,32; 0,75
0,41-0,45	27	34	(27/1028)/(34/365) = **0,28**	0,17; 0,46
0,36-0,40	13	37	(13/1028)/(37/365) = **0,12**	0,07; 0,23
0,31-0,35	8	44	(8/1028)/(44/365) = **0,06**	0,03; 0,14
≤ 0,30	19	182	(19/1028)/(182/365) = **0,04**	0,02; 0,06
Total	1028	365		

*IC: intervalo de confiança.

Fonte: Heffner.[5]

Quando o teste é classificado em vários níveis, a LR é calculada para cada nível, o que torna o teste com maior valor para aumentar ou reduzir a probabilidade da doença em relação ao teste classificado em dois níveis. A LR mostrada na Tabela 6 em cada nível da RPPS foi calculada dividindo a proporção de resultados da RPPS em condições com exsudato pela proporção de resultado da RPPS em condições com transudato. Como exemplo, a LR de 53,3 no nível 0,66-0,70 da RPPS foi determinada pela divisão da proporção 0,14591 (resultante de 150/1028) com a proporção 0,00274 (resultante de 1/365). Similarmente, a LR de 0,06 no nível 0,31-0,35 foi determinada pela divisão da proporção 0,0078 (resultante de 8/1028) com a proporção 0,1205 (resultante de 44/365).

Uso de múltiplos testes

No processo do diagnóstico de doenças, o clínico frequentemente solicita mais de um teste diagnóstico. Os testes são solicitados simultânea (testes paralelos) ou sequencialmente. Seguem exemplos de uso simultâneo e sequencial de testes diagnósticos.

Testes simultâneos (testes paralelos)

Os testes diagnósticos costumam ser solicitados simultaneamente quando a rápida avaliação do paciente é necessária. Para exemplificar o que é esperado ocorrer com o uso simultâneo de testes diagnósticos, imagine dois testes solicitados simultaneamente na avaliação diagnóstica de um paciente aqui referidos como teste A e teste B. Em seguida são mostradas a sensibilidade e a especificidade destes testes diagnósticos (Tabela 7) e a distribuição de valores de cada teste em uma amostra com prevalência de 20% (Tabela 8).

Tabela 7. Sensibilidade e Especificidade dos Testes Diagnósticos

	Teste A	Teste B
Sensibilidade	80%	90%
Especificidade	60%	90%

Fonte: Autores.

Tabela 8. Distribuição dos valores dos testes diagnósticos.

	Doença		
Teste A	Sim	Não	Total
Positivo	160	320	480
Negativo	40	480	520
Total	200	800	1000

Teste B	Sim	Não	Total
Positivo	180	80	260
Negativo	20	720	740
Total	200	800	1000

Fonte: Autores.

Quando os testes são usados simultaneamente, o resultado será positivo se for positivo para qualquer teste e negativo se negativo para ambos. Na Tabela 9 são mostrados os números de verdadeiros positivos e negativos resultantes do uso simultâneo do teste A e B. Como o teste A tem 0,8 (80%) de sensibilidade, ele identifica como positivo 144 do teste B, ou seja, 0,8 x 180. Número similar (144) é obtido usando a sensibilidade do teste B e positivos verdadeiros do teste A. Então 144 é o número de positivos tanto pelo teste A quanto pelo teste B.

Tabela 9. Verdadeiros positivos e negativos com o uso simultâneo dos testes.

Cálculo do novo número de verdadeiros positivos		Cálculo do novo número de verdadeiros negativos
Identificados como positivo por A e B	0,8 x 180 ou 0,9 x 160 = **144**	0,6 x 720 ou 0,9 x 480 = **432**
Identificado como positivo apenas por A	160 - 144 = **16**	
Identificado como positivo apenas por B	180 - 144 = **36**	
Total de positivos	144 + 16 + 36 = **196**	

Fonte: Autores.

Subtraindo 144 de 160 dos positivos verdadeiros pelo teste A, é obtido o número de positivos verdadeiros identificados apenas por A, ou seja, 16. De forma similar, subtraindo 144 de 180 dos positivos verdadeiros pelo teste B, é obtido o número de positivos verdadeiros identificados apenas por B, ou seja, 36. O total de positivos resultante do uso simultâneo é 196, ou seja, 144 + 16 + 36.

O número dos verdadeiros negativos com o uso simultâneo dos dois testes, 432, é obtido multiplicando a especificidade do teste A (0,6) pelo número de verdadeiros negativos do teste B (720) ou multiplicando a especificidade do teste B (0,9) pelo número de verdadeiros negativos do teste A (480). Na Tabela 10 são mostrados os números resultantes do uso simultâneo dos testes.

Tabela 10. Resultante do uso simultâneo dos testes.

	Doença		
Teste	Sim	Não	Total
Positivo	196	368	564
Negativo	4	432	436
Total	200	800	1000

Fonte: Autores.

Conforme mostrado na Tabela 11, o uso simultâneo dos testes resultou em aumento da sensibilidade e LR- de 0,04, portanto bem menor do que a LR- do teste A (0,33) e a LR- do teste B (0,11). Estes dados indicam aumento da capacidade de afastar a doença quando ambos os testes são negativos. No entanto, o uso de testes simultâneos tende a aumentar resultados falsos positivos e a prevalência aparente da doença. Conforme mostrado na Tabela 11, a prevalência aparente da doença com o uso simultâneo dos testes diagnósticos foi de 56%, muito superior à prevalência real de 20%.

Tabela 11. Medidas referentes ao teste A e ao teste B e resultante do uso combinado.

	Teste A	Teste B	Resultante do uso combinado
Prevalência aparente	48%	26%	**56%**
Prevalência real	20%	20%	20%
Sensibilidade	80%	90%	**98%**
Especificidade	60%	90%	54%
VPP	33%	69%	35%
VPN	92%	97%	99%
LR+	2,0	9,0	2,13
LR-	0,33	0,11	**0,04**

VPP: valor preditivo positivo; VPN: valor preditivo negativo.
Fonte: Autores.

Testes sequenciais

Quando testes são usados de forma sequencial, o teste subsequente é realizado quando o anterior é positivo, visando à confirmação. Imagine que 2.000 pacientes com um determinado problema são avaliados com suspeita de uma doença, usando inicialmente um teste com sensibilidade de 70% e especificidade de 80% em uma população clínica com prevalência da doença de 25%. Os resultados obtidos são mostrados na Tabela 12.

Tabela 12. Distribuição dos valores do teste inicial.

	Doença		
Teste	Sim	Não	Total
Positivo	350	300	**650**
Negativo	150	1200	1350
Total	500	1500	2000

Fonte: Autores.

Com uma sensibilidade de 70%, o teste identificaria 350 de 500 como verdadeiros positivos. Com uma especificidade de 80%, o teste identificaria 1.200 de 1.500 como verdadeiros negativos e 300 dos 1500 seriam

identificados como falsamente positivos. Dessa forma, 650 pacientes teriam teste positivo e seriam avaliados com o segundo teste.

Imagine que esse segundo teste tem 90% de sensibilidade e de especificidade. O total de indivíduos com a doença é 350. Como a sensibilidade é de 90%, 315 são positivos verdadeiros (Tabela 13). Já que 300 dos 650 não têm a doença e a especificidade é 90%, então 270 dos 300 serão verdadeiros negativos e 30 dos 300 serão falsamente negativos.

Tabela 13. Distribuição dos valores do segundo teste.

	Doença		
Teste	Sim	Não	Total
Positivo	315	30	345
Negativo	35	270	305
Total	350	300	**650**

Fonte: Autores.

As Tabelas 14 e 15 mostram o resultado final dos testes usados sequencialmente. Do total de 500 doentes, 315 foram positivos verdadeiros, correspondendo a uma sensibilidade de 63%. Portanto, ocorre uma redução da sensibilidade quando os testes são usados sequencialmente. Para calcular a especificidade resultante, observe que 1.200 foram negativos verdadeiros para o primeiro teste e não foram testados para o segundo teste. Um adicional de 270 pacientes foi negativo para o segundo teste. Ao se somar esses números, é obtido 1.470, que é o número resultante de negativos verdadeiros.

Tabela 14. Resultante do uso dos testes em série.

	Doença		
Teste	Sim	Não	Total
Positivo	315	30	345
Negativo	185	1470	1655
Total	500	1500	2000

Fonte: Autores.

Tabela 15. Medidas referentes ao testes A e B e resultante do uso sequencial.

	Teste A	Teste B	Resultante do uso combinado
Prevalência aparente	33%	53%	17%
Prevalência real	25%	54%	25%
Sensibilidade	70%	90%	63%
Especificidade	80%	90%	98%
VPP	54%	91%	91%
VPN	89%	89%	89%
LR+	3,50	9,0	31,5
LR-	0,38	0,11	0,38

VPP: valor preditivo positivo; VPN: valor preditivo negativo.

Fonte: Autores.

Conforme mostrado na Tabela 15, o uso sequencial dos testes resultou em aumento da especificidade e aumento da LR+, que significa aumento da capacidade de confirmar a doença.

Uso de testes na fase assintomática da doença

Este capítulo é voltado para o uso de estratégias visando ao diagnóstico de uma doença ainda na fase assintomática ou antes do aparecimento de achados clínicos, ou seja, rastreamento (*screening*). O objetivo de detectar a doença antes que ela se manifeste clinicamente é reduzir a morbidade e a mortalidade associadas com a doença. A premissa é que o tratamento iniciado mais precocemente é mais efetivo do que o tratamento iniciado mais tardiamente. Infelizmente, para várias situações, não existe tratamento que seja efetivo em reduzir a morbidade e mortalidade quando iniciado mais precocemente. Além disso, os riscos e custos relacionados com *screening* podem suplantar os potenciais benefícios da detecção mais precoce. Um dos malefícios é devido à carga psicológica de ser rotulado como uma doença que pode não se tornar sintomática, independentemente do tratamento. Os testes diagnósticos

apresentam resultados positivos falsos que geram uso desnecessário de testes diagnósticos e risco de efeitos adversos de tratamentos desnecessários. Portanto, é importante conhecer para quais doenças e em quais condições a detecção precoce da doença pode resultar em mais benefícios do que malefícios.

Abaixo estão questões para concluir se a identificação precoce de doenças é benéfica em determinada situação.

1. Existe evidência científica que permita concluir que o diagnóstico da doença antes do aparecimento de manifestações clínicas resulta em melhora da sobrevida e/ou qualidade de vida?
2. Existe teste diagnóstico adequado para confirmar e afastar o diagnóstico?
3. Existe tratamento efetivo para a doença e os pacientes diagnosticados precocemente aderem ao tratamento prescrito?
4. Os benefícios e malefícios variam na dependência de características dos pacientes, como idade e história familiar, de forma a justificar estratégias de screening com foco em grupos específicos?
5. A prevalência da doença e a morbidade/mortalidade associadas são elevadas e graves para justificar os custos da estratégia de *screening*?

Avaliação de estratégias de *screening*

A exequibilidade e a efetividade são pontos básicos na avaliação de um programa de *screening*. A exequibilidade depende da aceitabilidade, custo-efetividade e disponibilidade de testes para avaliações diagnósticas subsequentes. Para ser aceita pelas pessoas, a estratégia de *screening* deve ser simples, feita de forma rápida e sem desconforto. O exame Papanicolau para detecção precoce do câncer do colo uterino, por exemplo, é rápido, indolor e bem aceito pelas mulheres. Em contraste, a sigmoidoscopia para o câncer colorretal é um exame demorado e desconfortável. Em virtude dessa característica, é pequena a proporção da população que é submetida à sigmoidoscopia na faixa etária para a qual o exame tem sido indicado.[6] A exequibilidade de *screening* para

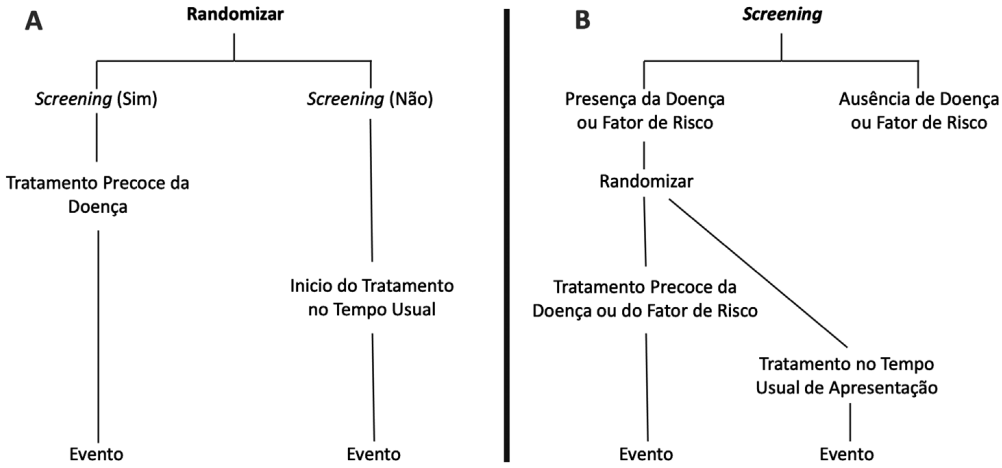
determinadas doenças deve também considerar a disponibilidade de seguimento para reavaliações clínicas, testes diagnósticos adicionais e tratamento.

A efetividade de uma estratégia de *screening* é medida, principalmente, pela redução da morbidade e mortalidade. Mesmo que uma estratégia de *screening* mostre elevada exequibilidade e seja capaz de identificar de forma acurada e com baixo custo grande número de indivíduos, na fase assintomática da doença, ou antes que apareçam as manifestações clínicas, ela terá pouco impacto nos contextos da clínica e da saúde pública se o diagnóstico e o tratamento precoces não resultarem na redução da morbidade e mortalidade. Estudos observacionais, como o estudo de coorte[7] e estudo de caso-controle,[8] têm sido usados na avaliação de efetividade de estratégias de *screening*. No entanto, devido à suscetibilidade dos estudos observacionais para vieses, os ensaios clínicos randomizados são os que geralmente propiciam o nível de evidência mais elevado sobre a efetividade de estratégias de *screening*.[9,10]

A Figura 2 mostra dois tipos de desenho de ensaio clínico randomizado para testar a efetividade de uma estratégia de *screening*.[11] No desenho mostrado na Figura 2A, os participantes são inicialmente randomizados para serem ou não submetidos à estratégia de *screening*. Este desenho avalia a efetividade do *screening* em detectar a doença na fase assintomática e do início mais precoce do tratamento da doença. Este desenho tem sido usado em estudos de efetividade de *screening* para câncer, como câncer colorretal, prostático e da mama.[12]

No desenho mostrado na Figura 2B, todos os participantes recebem o *screening*. Os que tiverem doença diagnosticada na fase assintomática, ou fator de risco detectado, são então randomizados para iniciar tratamento imediato ou iniciar tratamento no tempo usual. Se os participantes que recebem o tratamento evoluem melhor, então é concluído que o tratamento precoce é benéfico. Esse tipo de desenho foi usado em teste de efetividade de *screening* para detecção de hipertensão arterial e níveis elevados de colesterol.[13,14]

Figura 2. Desenhos de estudo para avaliar efetividade de estratégia de *screening*.



Fonte: modificado de Barratt.[11]

Vieses em estudos de efetividade de *screening*

Vieses frequentes em estudos de efetividade de estratégias de *screening* são o viés do voluntário ou viés de autosseleção, viés do tempo ganho (*lead time bias*), viés do tempo de duração (*length time bias*) e o viés de sobrediagnóstico (*overdiagnosis*).

O viés do voluntário é devido ao fato que voluntários para participar de programa de *screening* são, em geral, diferentes no estado de saúde em relação aos que não participam do programa. Em estudos observacionais, viés pode distorcer os resultados se a proporção de voluntários é diferente entre o grupo submetido e o grupo não submetido à estratégia de rastreamento. O viés do voluntário não se constitui um problema em ensaio clínico randomizado porque a alocação para os grupos de *screening* ocorre após os participantes concordarem em participar do estudo.

O viés do tempo ganho (*lead time bias*) ocorre ao se concluir erroneamente que o *screening* está associado com maior sobrevida sem considerar que o diagnóstico da doença foi feito ainda na fase pré-clínica. É necessário considerar o tempo ganho devido ao diagnóstico ter sido feito na fase pré-clínica no grupo submetido ao *screening* para evitar a conclusão errônea de que a estratégia de *screening* resultou em melhora da sobrevida.[15]

Viés do tempo de duração (*length time bias*) ocorre em estudos observacionais pela maior probabilidade de detectar pelo *screening* pessoas com fase pré-clínica mais longa que, em geral, têm melhor prognóstico.

Viés de sobrediagnóstico (*overdiagnosis*) ocorre quando o programa de rastreamento identifica uma doença que irá permanecer até a morte sem se manifestar clinicamente. O sobrediagnóstico é considerado uma situação extrema de viés do tempo de duração e é frequente em alguns tipos de câncer, como o câncer de próstata e de mama.

Número necessário para rastrear (NNR)

Representa o número de pessoas que precisam ser submetidas à estratégia de *screening* para prevenir um caso de morte ou outro evento de saúde.[16] O cálculo é similar ao número necessário para tratar (NNT).

$$NNR = \frac{1}{\text{redução absoluta do risco em proporção}}$$

ou

$$NNR = \frac{100}{\text{redução absoluta do risco em percentual}}$$

Reprodutibilidade do teste

Na avaliação de um teste diagnóstico, é importante saber o quanto esse teste é reprodutível, o que significa a concordância do mesmo resultado por diferentes examinadores ou pelo mesmo examinador em diferentes momentos. O índice Kappa é uma medida usada para avaliar reprodutibilidade. Essa medida representa a proporção do máximo de concordância possível de ser obtido além da concordância observada. Um índice Kappa de 0,41-0,60 é considerado concordância moderada; 0,61-0,80, concordância substancial; e 0,81-1, concordância quase perfeita.[17] Abaixo é mostrada a fórmula do índice Kappa.

$$Kappa = \frac{(\text{proporção de concordância observada-proporção de concordância por chance})}{(1 - \text{proporção de concordância por chance})}$$

Tabela 16. Dados hipotéticos da concordância entre dois clínicos ao determinar o resultado de um teste diagnóstico.

A. Números e proporção de concordância observada				
	Segundo examinador			Proporção de concordância observada
Primeiro examinador	**Teste positivo**	**Teste negativo**	**Total**	
Teste positivo	**100**	20	120	$\frac{(100 + 60)}{200} = \mathbf{0,8}$
Teste negativo	20	**60**	80	
Total	120	80	200	

B. Números e proporção de concordância por chance				
	Segundo examinador			Proporção de concordância por chance
Primeiro examinador	**Teste positivo**	**Teste negativo**	**Total**	
Teste positivo	$\frac{(120 \times 120)}{200} = \mathbf{72}$	$\frac{(80 \times 120)}{200} = 48$	120	$\frac{(72 + 32)}{200} = \mathbf{0,52}$
Teste negativo	$\frac{(120 \times 80)}{200} = 48$	$\frac{(80 \times 80)}{200} = \mathbf{32}$	80	
Total	120	80	200	

Fonte: Autores.

A Tabela 16 mostra dados hipotéticos de dois examinadores que avaliaram um teste diagnóstico classificado como positivo e negativo em uma amostra de 200 pacientes. Ambos classificaram o teste como positivo em 100 pacientes e como negativo em 60, correspondendo a uma concordância observada de 0,8 (100+60)/200. Os números de concordantes por chance foram determinados para cada célula multiplicando o total de cada coluna pelo total de cada linha e dividindo pelo número total de pacientes. A proporção de concordância por chance foi 0,52 (72+32)/200. Usando a concordância observada e a concordância por chance na fórmula mostrada anteriormente, foi calculado o índice Kappa como 0,58 (0,8-0,52)/(1-0,52), o que indica uma concordância moderada entre os dois examinadores.

Referências

1. Light RW, Macgregor MI, Luchsinger PC, Ball WC Jr. Pleural effusions: the diagnostic separation of transudates and exudates. An Inter Med. 1972; 77(4): 507-13.

2. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lancet. 2005; 365(9469): 1500-5.

3. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anemia: an overview. J Gen Intern Med. 1992; 7: 145-53.

4. Morgan TO, Jacobsen SJ, McCarthy WF, Jacobson DJ, McLeod DG, Moul JW. Age-specific reference ranges for serum prostate-specific antigen in black men. N Engl J Med. 1996; 335(5): 304-10.

5. Heffner JE, Sahn SA, Brown LK. Multilevel likelihood ratios for identifying exudative pleural effusions. Chest. 2002; 121: 1916-20.

6. Schroy PC, 3rd, Wilson S, Afdhal N. Feasibility of high-volume screening sigmoidoscopy using a flexible fiberoptic endoscope and a disposable sheath system. Am J Gastroenterol. 1996; 91(7): 1331-7.

7. Hellquist BN, Duffy SW, Abdsaleh S, Björneld L, Bordás P, Tabár L, et al. Effectiveness of population-based service screening with mammography for women ages 40 to 49 years: evaluation of the Swedish Mammography Screening in Young Women (SCRY) cohort. Cancer. 2011; 117(4): 714-22.

8. Maroni R, Massat NJ, Parmar D, Dibden A, Cuzick J, Sasieni PD, et al. A case-control study to evaluate the impact of the breast screening programme on mortality in England. Br J Cancer. 2020; 124: 736-43.

9. Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, et al. Prostate-cancer mortality at 11 years of follow-up. N Engl J Med. 2012; 366(11): 981-90.

10. Shaukat A, Mongin SJ, Geisser MS, Lederle FA, Bond JH, Mandel JS, et al. Long-term mortality after screening for colorectal cancer. N Engl J Med. 2013; 369(12): 1106-14.

11. Barratt A, Irwig L, Glasziou P, Cumming RG, Raffle A, Hicks N, et al. Users' guides to the medical literature: XVII. How to use guidelines and recommendations about screening. Evidence-Based Medicine Working Group. JAMA. 1999; 281(21): 2029-34.

12. Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. BMJ. 1988; 297(6654): 943-8.

13. Multiple risk factor intervention trial. Risk factor changes and mortality results. Multiple Risk Factor Intervention Trial Research Group. JAMA. 1982; 248(12): 1465-77.

14. Frick MH, Elo O, Haapa K, Heinonen OP, Heinsalmi P, Helo P, et al. Helsinki Heart Study: primary-prevention trial with gemfibrozil in middle-aged men with dyslipidemia. Safety of treatment, changes in risk factors, and incidence of coronary heart disease. N Engl J Med. 1987; 317(20): 1237-45.

15. Baker SG, Kramer BS, Prorok PC. Statistical issues in randomized trials of cancer screening. BMC Med Res Methodol. 2002; 2: 11.

16. Rembold CM. Number needed to screen: development of a statistic for disease screening. BMJ. 1998; 317(7154): 307-12.

17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1): 159-74.