Atualização

Testes diagnósticos 2: as propriedades do teste

Diagnostic tests 2: the test properties

Denise Duprat Neves [1], *Ricardo Marques Dias* [1],
Antônio José Ledo Alves da Cunha [2]

Descritores: *diagnóstico, sensibilidade, especificidade, valor preditivo, curva ROC.*
Keywords: *diagnosis, sensibility and specificity , predictive value, ROC curve.*

Introdução

Para avaliar a utilidade de novos testes devemos questionar: 1) se os resultados são provenientes de um estudo válido; 2) quais são as propriedades do teste; e 3) quão úteis serão os resultados na prática clínica. No primeiro artigo desta série discutimos a primeira pergunta acima, relacionada à validade dos estudos de teste diagnostico. Se os resultados foram obtidos por meio de um estudo válido, o passo seguinte será avaliar o poder discriminatório do teste, ou seja, qual o seu desempenho na identificação da doença. A acurácia de um teste é geralmente descrita através da sensibilidade e especificidade, consideradas características do teste diagnóstico. Neste artigo vamos descrever e discutir a aplicação destas e de outras propriedades habitualmente utilizadas em estudos de testes diagnósticos.

Propriedades dos testes diagnósticos

As propriedades de testes diagnósticos são, geralmente, calculados a partir da distribuição de freqüência em tabela de contingência 2X2, conforme descrito na tabela 1. O diagnóstico deve ser estabelecido por um padrão de referência, também conhecido como padrão ouro, adequado e realizado de forma independente do teste em estudo [1, 2, 3, 4, 5, 6, 7, 8].

Tabela 1 – *Tabela de contingência para cálculo das propriedades dos testes diagnósticos.*

Resultado do teste	Diagnóstico	
	doente	não doente
positivo	**a**	**b**
negativo	**c**	**d**

Onde:

a = verdadeiro positivo
b = falsos positivos
c = falsos negativos
d = verdadeiro negativo
a+b = testes positivos
c+d = testes negativos
a+c = doença presente
b+d = doença ausente

1. A *sensibilidade* (S) e *especificidade* (E) são características do teste diagnóstico e valorizadas quando se solicita um teste ou quando estamos avaliando seu

1. Hospital Universitário Gaffrée e Guinle - UNI-RIO
2. Instituto de Puericultura Mastargão Gesteira - Universidade Federal do Rio de Janeiro

Endereço para correspondência: Denise Duprat Neves. Rua Mariz e Barros 775, Hospital Universitário Gaffrée e Guinle, DEMESP. Pneumologia. Tijuca, Rio de Janeiro. Brasil. CEP 20270-004. E-mail: dduprat@unirio.br. Tel: 55 021 2569 7610 – ramal 304.
Artigo recebido para publicação em 03/9/2003 e aceito no dia 13/10/2003, após revisão.

desempenho. A sensibilidade, S = (a / a+c) X 100, calcula a probabilidade de um teste dar um resultado positivo quando a pessoa submetida a ele é verdadeiramente doente. A especificidade, E = (d / b+d) X 100, expressa a probabilidade de um teste apresentar um resultado negativo quando a pessoa examinada não está doente. A partir destes dois valores poderemos calcular os demais, caso não venham descritos no estudo.

Testes com alta sensibilidade, caso negativo, ajudam a excluir, "descartar", o diagnóstico. São úteis na investigação inicial, busca de casos, principalmente de doenças graves ou tratáveis. Como exemplo, podemos citar o teste de ELISA anti-HIV realizado nos doadores de sangue. Por outro lado, quando é importante não ter resultados falso-positivos, daremos ênfase a uma alta especificidade, que são testes úteis para confirmar um diagnóstico caso seja positivo. Como exemplo de testes específicos podemos citar a cultura para pesquisa de *M. tuberculosis* e o exame histopatológico identificando células malignas.

Devemos destacar que não há possibilidade de um teste diagnóstico apresentar sensibilidade e especificidade de 100% ao mesmo tempo, pois estas propriedades variam inversamente dentro de uma determinada faixa de valores.

2. A *acurácia* (A) representa a utilidade do teste como um todo, expressando os resultados verdadeiros na amostra, e sendo calculada por: A = (a+d / a+b+c+d) X 100. É útil na comparação do rendimento de testes, pois resume a um número os resultados verdadeiros obtidos com o teste. No entanto, não é ideal para escolha ou utilização de testes diagnósticos, uma vez que não individualiza as características do teste, a sensibilidade e a especificidade. Uma acurácia de 90% pode ser obtida com uma sensibilidade de 97% e especificidade de 83% ou vice-versa.

3. *Valor preditivo* é uma probabilidade condicional que expressa a probabilidade da doença estar ou não presente, em função de determinado resultado. É a maneira pela qual utilizamos os testes diagnósticos na prática. O valor preditivo positivo (VPP) = (a / a+b) X 100, permite estimar a probabilidade de um indivíduo ser verdadeiramente doente, uma vez que o resultado do teste tenha sido positivo. O valor preditivo negativo (VPN) = (d / c+d) X 100, revela a probabilidade do indivíduo não estar doente, dado que seu teste tenha tido um resultado negativo.

Estamos agora "olhando" a tabela de contingência pela direção dos resultados. Assim, estes valores dependem não só da sensibilidade e especificidade, mas também da prevalência da doença na amostra. Quanto mais rara a doença, maior a certeza de que um teste negativo indique a sua ausência, e menor a probabilidade de um teste positivo indicar sua presença (tabela 2).

Tabela 2 – *Mesmo teste aplicado em populações com diferente prevalência de doença.*

Teste	Doença	
Diagnóstico	+	-
+	19	99
-	1	1881

Teste	Doença	
Diagnóstico	+	-
+	57	2
-	3	38

Vamos supor um teste com sensibilidade e especificidade iguais a 95%. Aplicando este teste para busca de casos em pessoas assintomática, digamos em um shopping ou numa praça, a prevalência de doença será menor do que 1%. Assim teremos um VPP=16,1% e VPN de 99,9%. Agora vamos aplicar este mesmo teste em pacientes de um ambulatório de referência, onde buscamos confirmar o diagnóstico. Neste grupo existe uma maior probabilidade dos indivíduos testados estarem doentes, digamos que uma prevalência de 60%. Neste grupo o mesmo teste apresenta um VPP de 96,6% e o VPN de 92,7%. Sendo assim, os resultados obtidos em um estudo não devem ser utilizados genericamente. A prevalência da doença pode variar em função da região geográfica de atuação ou mesmo do local de atendimento, se na rede de atendimento primária ou em um hospital de referência.

4. A *Prevalência* (Prev) mede a freqüência da doença na população estudada, sendo calculada por: Prev = (a+c / a+b+c+d) X 100. A prevalência da doença, associada a dados da história e exame físico, pode influenciar a nossa crença na presença de determinada doença e, em algumas situações, pode ser utilizada como a probabilidade pré-teste de sua presença. Quando conhecemos as características do teste, podemos calcular os valores preditivos para qualquer prevalência, através das seguintes fórmulas, baseadas no teorema de Bayes: VPP = S*Prev / S*Prev + (1-E) * (1-Prev) e VPN = E*(1-Prev) / (1-S) * Prev + E*(1-Prev).

5. As *Razões de verossimilhança* (LR, do inglês *"Likelihood Ratio"*) são também chamadas de razão ou índice de probabilidade. É uma razão, de duas proporções, que expressa quantas vezes determinado resultado, positivo ou negativo, surge mais nos doentes do que nos não doentes. As fórmulas para seu cálculo são: da razão de verossimilhança positiva (LR+) = {a/(a+c)} / {b/(b+d)} ou S / (1–E), expressando a probabilidade de teste positivo nos doentes em razão do mesmo resultado nos não doentes, e a da razão de

verossimilhança negativa (LR-) = {c/(a+c)} / {d/(b+d)} ou (1-S) / E. Estas têm se tornado a maneira preferida de se expressar a utilidade e de se comparar testes diagnósticos.

Uma LR maior do que 1 irá aumentar a probabilidade da presença da doença, assim como uma LR menor do que este valor diminui esta chance. Quando a probabilidade pré-teste se encontra entre 30 a 70%, testes com LR+ maior do que 10 praticamente confirmam a presença de doença. Uma razão de verossimilhança positiva de 50, por exemplo, significa que a chance da doença estar presente é 50 vezes maior diante de um resultado positivo do que a chance antes de conhecermos o resultado do teste. Seu uso possuiu a vantagem de poder ser utilizado seqüencialmente, sendo que a probabilidade pós-teste após a utilização de um teste passa a ser utilizada como probabilidade pré-teste de outros testes. Pode ser calculada, ainda, para diferentes valores discriminatórios, no caso de testes com resultado em valores contínuos. Sua utilização na prática será discutida em maiores detalhes no próximo artigo desta série.

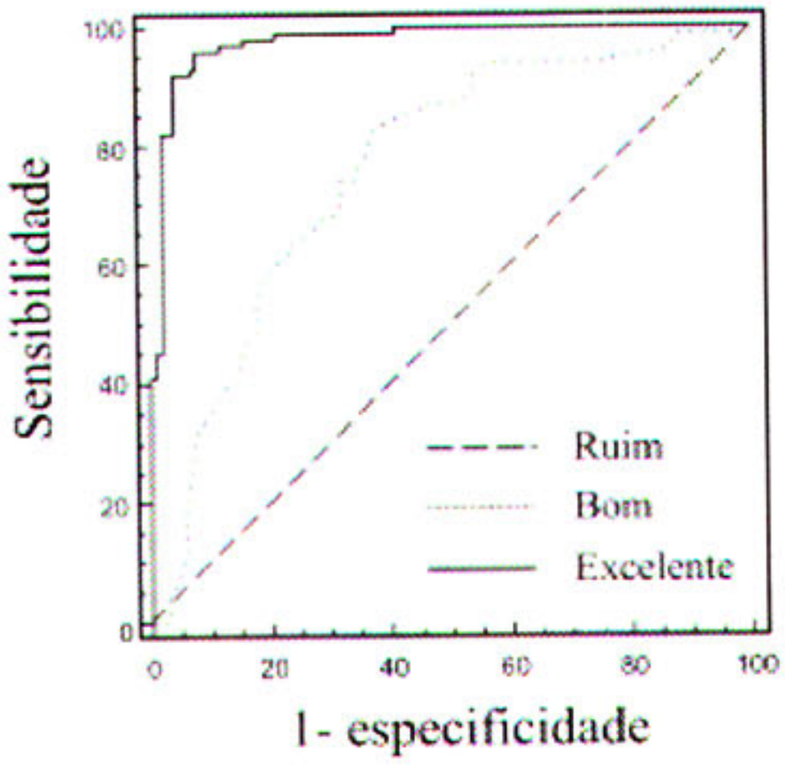
A curva ROC

Quando realizamos um teste diagnóstico queremos saber sobre a presença ou ausência de doença, uma resposta dicotômica. No entanto, muitos testes têm como resultado valores contínuos. É necessário, portanto, determinar um valor discriminatório que distingue o normal do anormal. Esta não é uma medida simples e para alguns testes existe uma faixa de normalidade e para outros uma faixa indeterminada, conhecida como "zona cinza". A curva ROC ("*receiver operating characteristic*"), desenvolvida nos anos 50 para avaliação de sinais de radar, tem sido utilizada, nas últimas duas décadas, para determinar o melhor valor discriminatório em testes diagnósticos.

A curva ROC é construída pelos pontos que correspondem a sensibilidade, no eixo do "X" ou vertical, e 1-especificidade, no eixo do "Y" ou horizontal, para cada valor observado na amostra [9], figura 1. A tabela gerada deve conter no mínimo cinco pontos diferentes de valores discriminatórios, sendo preferível, além do valor que permite obter a maior acurácia (menor número de falsos resultados), a descrição daqueles que correspondem a sensibilidade e a especificidade de 90%, 95% e 99% [4]. Permite, assim, escolher o valor discriminatório na dependência da necessidade clínica. Se o teste for utilizado na busca de casos entre assintomáticos ("*screening*") devemos escolher um valor que permita uma alta sensibilidade. Os casos positivos deverão ser confirmados por outro teste. Este, por sua vez, deverá ter uma alta especificidade [10].

A tendência da sensibilidade e especificidade variarem inversamente dentro de uma determinada faixa de valores dificulta, não só, a tarefa de se determinar o melhor valor discriminatório, mas também a de se comparar o desempenho dos testes diagnósticos. A medida da área abaixo da curva (AUC) ROC pode ser útil neste aspecto. Testes com bom poder discriminatório apresentam curvas que se aproximam do eixo vertical, demonstrando que à medida que a sensibilidade aumenta, existe pouca ou nenhuma perda da especificidade. O teste perfeito, tem uma área igual à unidade. Curvas podem ser comparadas em relação à linha diagonal, bissetriz entre os eixos, que representa um teste não capaz de diferenciar os dois grupos. Quando a variável em estudo tem este comportamento, a área abaixo da curva se aproxima de 0,5. Como regra geral, de maneira acadêmica, o seguinte sistema de classificação foi proposto: excelente se entre 1 e 0,90, bom de 0,90 a 0,80, regular de 0,80 a 0,70, ruim de 0,70 a 0,60 e falho quando entre 0,60 e 0,50.

Figura 1 – *A curva ROC e comparação entre testes.*



A comparação entre dois ou mais testes pode ser visual (figura 1), mas existem testes estatísticos para se comparar a área abaixo da curva, para dados pareados ou não, utilizando testes paramétricos ou não [11, 12].

Como escolher um teste diagnóstico?

Como visto, a valorização de maior sensibilidade ou especificidade depende do contexto para o qual o teste será aplicado. Testes com sensibilidade de 95% e especificidade de 80% podem ser adequados para diagnosticar pacientes daltônicos, por exemplo. Por outro lado, testes para busca de casos de doenças com

graves conseqüências, como o "teste do pezinho" para o diagnóstico de hipotireoidismo congênito ou o teste para HIV nos doadores de sangue, devem ser altamente sensíveis. Estes testes possuem, em contrapartida, um baixo valor preditivo positivo, necessitando de um outro teste que confirme a presença da doença.

O progresso e a tecnologia dos novos testes diagnósticos vem valorizando o poder de identificar doenças, tornado os testes mais sensíveis. Por outro lado, o que está cada vez mais raro é a palavra "patognomônico", ou seja, característico de determinada doença, a distinguindo de outras. Individualmente, para o paciente, e coletivamente, para os serviços de saúde, a especificidade é provavelmente mais importante.

Compreensão destes parâmetros na prática clínica

Um outro aspecto a ser avaliado é o como os médicos recebem e utilizam esta informação. Geralmente existe uma dificuldade em transferir para a prática os resultados de estudos científicos. Pesquisa realizada com 300 médicos americanos, de 50 especialidades, sobre como utilizam testes diagnósticos, mostrou que poucos conhecem ou utilizam a metodologia Bayesiana (3%), curva ROC ou LRs (1% cada). Embora mais de 80% afirmassem que utilizam a sensibilidade e especificidade em algum momento da decisão clínica ou na avaliação de novos métodos, o faziam de modo informal [13].

Cabe ao pesquisador traduzir, tentar fazer sentido, como um determinado teste "funciona" em determinada população, e explicitar o contexto ao qual ele se aplica. Médicos preferem algorítmicos simples, especialmente quando disponíveis em computadores, do que entender, interpretar e manusear os parâmetros descritos (S, E, VPs e LRs).

Devido a dificuldade de compreensão e aplicação pelos médicos dos resultados obtidos em estudos sobre testes diagnósticos, vários procedimentos tentam simplificar a utilização destes. É o que será discutido no próximo artigo desta série.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Fletcher RH, Fletcher SW, Wagner EH. Epidemiologia Clínica: elementos essenciais. 3 ed. Porto Alegre: Artes Médicas; 1996.
2. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What were the results and will they help me in caring for my patients? JAMA 1994; 271(9):703-7.
3. Greenhalgh T. How to read a paper: Papers that report diagnostic or screening tests. Br Med J 1997; 315:540-543.
4. Altman DG, Bland JM. Statistics Notes: Diagnostic tests 1: sensitivity and specificity. Br Med J 1994;308:1552.
5. Gambino R. The misuse of predictive value - or why you must consider the odds. Ann Ist Super Sanita 1991; 27(3):395-9.
6. Altman DG, Bland M. Diagnostic tests 2: predictive values. Br Med J 1994;309:102.
7. Medronho RA, Carvalho DM, Bloch KV, Luiz RR, Werneck GL. Epidemiologia. São Paulo: Atheneu; 2002.
8. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Evidence base of clinical diagnosis: Designing studies to ensure that estimates of test accuracy are transferable. Br Med J 2002;324:669-71.
9. Altman DG, Bland M. Diagnostic tests 3: receiver operating characteristic plots. Br Med J 1994; 309:188.
10. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Ann Inter Med 1981; 94:555-600.
11. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993; 39:561-77.
12. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 1988; 44:837-45.
13. Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgements: practicing physicians' use of quantitative measures of test accuracy. Am J Med 1998; 104:374-80. ■