Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Avaliação de Testes Diagnósticos

Edna Afonso Reis
Ilka Afonso Reis

Relatório Técnico
RTP-02/2002
Série Ensino

ÍNDICE

1. Introdução

O diagnóstico de uma doença é um procedimento delicado por involver a vida de "pacientes", sejam eles seres humanos, animais e até mesmo plantas. Como instrumento auxiliar no processo de diagnóstico, é comum o uso de testes construídos a partir de medições feitas nos pacientes, sejam elas quantitativas ou qualititativas. Esses testes são chamados *testes clínicos* ou *testes diagnósticos*.

Atualmente, esses instrumentos tornaram-se tão importantes e difundidos que existem grandes indústrias e laboratórios inteiramente dedicados à produção de testes diagnósticos cada vez mais precisos, mais rápidos e, às vezes, mais baratos.

Sendo os testes diagnósticos procedimentos humanos, estão sujeitos a erro. Desse modo, antes que um teste seja adotado como instrumento auxiliar no diagnóstico de determinada doença, é necessário que suas possibilidades de erro sejam avaliadas. As medidas de avaliação da qualidade de um teste clínico e do diagnóstico baseado nele são o assunto deste texto.

Quando falamos em diagnóstico, não pensamos apenas em exames ou testes de detecção de doenças. Podemos imaginar todo tipo de recurso para identificar uma dentre duas situações, como o detector de mentira que, baseado em alterações da voz, dos batimentos cardíacos ou respiração, tem o objetivo de identificar uma mentira ou verdade. Ou ainda um exame *anti-dopping* em atletas, onde a "doença" seria o fato do atleta ter ingerido algum tipo de droga. Porém, como a grande maioria dos testes clínicos está relacionada ao diagnóstico de doenças, essa abordagem tornou-se a mais comum. Neste texto, usaremos os termos "doença", "doente" e "sadio", sem nos esquecer de que as idéias aqui apresentadas podem ser aplicadas a outros contextos, como mostramos em alguns dos exemplos.

Antes de conhecermos as medidas de qualidade para um teste clínico e para o diagnóstico baseado nele, devemos estudar um pouco de Probabilidade, que nos ajudará a lidar com a incerteza do diagnótico a ser dado ao paciente.

2. Noções de Probabilidade

Probabilidade pode ser definida como a medida da (in)certeza sobre o acontecimento de um fato ou evento. Em nossa vida diária estamos freqüentemente tomando decisões baseadas na probabilidade, mesmo que, muitas vezes, não nos demos conta disso. Por exemplo, se, ao sairmos de casa pela manhã, não estamos certos sobre o fato de o tempo continuar ensolarado, procuramos nos precaver levando um guarda-chuva. Inconscientemente, atribuímos um grau de incerteza à ocorrência de chuva baseados nas nuvens no céu, umidade no ar, comportamento dos bichos, etc.

Em algumas situações, é possível calcular a probabilidade de ocorrência de um evento. Vamos começar a estudar os conceitos básicos de probabilidade nessas situações.

2.1. Conceitos Básicos

Vamos definir *experimento aleatório*, *espaço amostral* e *evento*, conceitos importantes na definição de probabilidade.

Experimento aleatório é aquele no qual os resultados possíveis são conhecidos, mas não sabemos de antemão qual deles irá ocorrer.

São exemplos de experimentos aleatórios:

Experimento aleatório 1: retirar uma carta de um baralho de 52 cartas e verificar sua cor;
Experimento aleatório 2: retirar uma carta de um baralho de 52 cartas e verificar o seu naipe;
Experimento aleatório 3: jogar um dado e observar a face de cima;
Experimento aleatório 4: sortear um estudante de uma lista e verificar a cor de seus olhos;
Experimento aleatório 5: sortear uma mulher na cidade e verificar quantos filhos ela tem.

O conjunto de resultados possíveis para um experimento é chamado de seu **espaço amostral,** (representado peloa letra E). Vejamos:

Espaço amostral do experimento aleatório 1: E_1 = {preta, vermelha};

Espaço amostral do experimento aleatório 2: E_2 = {copas, paus, ouros, espada};
Espaço amostral do experimento aleatório 3: E_3 = {1, 2, 3, 4, 5, 6};
Espaço amostral do experimento aleatório 4: E_4 = {preto, castanho, verde, azul};
Espaço amostral do experimento aleatório 5: E_5 = {0, 1, 2, 3, 4, 5, 6, 7, 8,....}.

Todo subconjunto do espaço amostral de um experimento aleatório é chamado **evento**.

No experimento aleatório 1, se definirmos o evento A "carta preta", ele será representado pelo conjunto A={preta}. O evento A é um exemplo de *evento simples*, pois contém somente um dos elementos do espaço amostral.

No experimento aleatório 2, podemos definir o evento B como sendo "carta vermelha". Assim, B = {copas,ouros}. Para o Experimento Aleatório 5, o evento F "mulher com número de filhos inferior a três" é representado pelo conjunto F = {0,1,2}.

No experimento aleatório 3, o evento D "face de cima igual a 7" é um exemplo de *evento vazio*, pois o dado não tem a face 7. O evento D é representado pelo conjunto vazio D = { } ou D = ∅.

Existe ainda um outro tipo de evento chamado *evento complementar*. O complementar de um evento é formado pelos elementos do espaço amostral que não fazem parte do evento. Por exemplo, considerando o evento B = {copas,ouros}, definido no espaço amostral do Experimento Aleatório 2, o complementar de B é formado por $\overline{B}$ = {paus, espada} . Juntos, os eventos B e $\overline{B}$ formam o espaço amostral E2. Daí o nome evento complementar.

Novos eventos sçao obtidos pela combinção de outros, atevés da "união" ou "interseção":

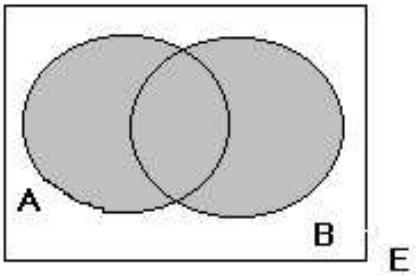
União de Eventos (A∪B): o evento "A união B" é formado pelos elementos que estão em A ou em B.

Interseção de Eventos (A∩B): o evento " A interseção B" é formado pelos elementos que estão em A e em B ao mesmo tempo.

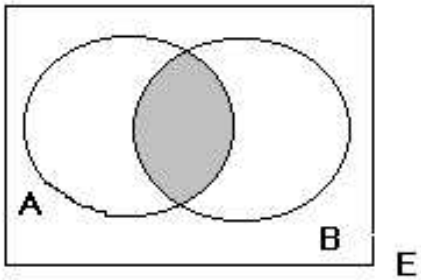
Exemplo 2.1: Sejam os eventos A "*sair um número par na jogada de um dado*" , A = { 2, 4, 6}
e B "*sair um número menor do que 4 na jogada de um dado*", B = {1, 2, 3} .
O evento união é formado por (A ∪ B) = {1, 2, 3, 4, 6} e o evento interseção é formado por (A ∩ B) = { 2 }.

O Diagrama de Venn é uma representação gráfica útil para a visualização das combinações de eventos. No diagramas abaixo os eventos resultantes da união e da interseção dos eventos A e B são representados pela área hachurada.

Evento (A ∪ B):



Evento (A ∩ B):



Dois eventos são ***mutuamente exclusivos*** quando não possuem interseção. Este nome vem do fato de que, se os eventos A e B não possuem elementos em comum, a ocorrência de um deles automaticamente exclui a ocorrência do outro e vice-versa. Por exemplo, considere o experimento aleatório "sortear cinco letras do alfabeto e formar uma palavra", cujo o espaço amostral é formado pelas 26 letras do alfabeto. Seja o evento A "as cinco primeiras letras do alfabeto", A = {a,b,c,d,e}, e o evento B "a palavra formada é POMBO". Se o experimento é realizado e o evento A ocorre, a ocorrência do evento B já está automaticamente excluída. Caso o evento B ocorra, o evento A estará automaticamente excluído. Um evento e seu complementar são sempre mutuamente exclusivos.

2.2. Definição de Probabilidade

Quando os elementos do espaço amostral têm a mesma probabilidade de ocorrerem, a *definição clássica da probabilidade* de um evento A ocorrer é dada por

$$P(A) = \frac{\text{número de elementos do evento A}}{\text{total de elementos do espaço amostral}}.$$

Exemplo 2.2: seja o experimento aleatório "jogar um dado e observar a face de cima". Se o dado for honesto, todas as seis faces têm a mesma probabilidade de saírem para cima. Assim, a probabilidade do evento F "o número é par", F = {2,4,6}, é dada por P(F)=3/6=0,5.

No entanto, a definição clássica não pode ser aplicada a situações onde os elementos do espaço amostral não possuam a mesma probabilidade de ocorrer, o que é o caso da maioria das situações práticas. Nesses casos, usaremos a *definição frequentista de probabilidade*, que fornece uma estimativa da probabilidade do evento de interesse através de sua frequência em um grande número de realizações independentes do experimento.

Exemplo 2.3: suponha que desejássemos conhecer a probabilidade de uma pessoa estar infectada com a bactéria *H. pylori* . O experimento aleatório consiste em selecionar uma pessoa do grupo de interesse e verificar um dos resultados possíveis: está infectada ou não está infectada. Estes dois eventos do espaço amostral não têm a mesma probabilidade de ocorrência, o que impossibilita o uso da definição clássica de probabilidade. Através do exame de uma grande amostra de pessoas, usaremos a definição frequentista e estimaremos a probabilidade de uma pessoa estar infectada com a *H. pylori* usando a frequência relativa de pessoas infectadas nessa amostra.

2.3. Propriedades da Probabilidade

Seja um experimento aleatório cujo o espaço amostral é dado pelo conjunto E. O evento A é um evento definido dentro do espaço amostral E. Para que um número possa ser considerado uma probabilidade, ele deve atender às seguintes propriedades.

1. $0 \leq P(A) \leq 1$
2. $P(E) = 1$
3. $P(\overline{A}) = 1 - P(A)$

A propriedade 1 diz que não existe probabilidade que seja negativa e nem maior do que 1. Ou seja, um evento pode ser impossível (probabilidade igual a zero), ou ocorrer com uma certa probabilidade ou acontecer com certeza (probabilidade igual a 1).

A propriedade 2 diz que a probabilidade de ocorra algum evento do espaço amostral é 1, ou seja, o resultado do experimento aleatório tem que ser um ou mais elementos do espaço amostral.

A propriedade 3 é muito útil, pois freqüentemente não sabemos como calcular a probabilidade do evento A diretamente, mas sabemos como calcular a probabilidade do seu complementar. Pela propriedade 3, podemos calcular a probabilidade de A indiretamente.

A **probabilidade da união de dois**[1] **eventos** é calculada pela seguinte equação:

$$\mathbf{P(A \cup B) = P(A) + P(B) - P(A \cap B).}$$

Quando os eventos A e B forem mutuamente exclusivos, $P(A \cup B) = P(A) + P(B)$, pois $P(A \cap B) = 0$.

1 A expressão para o cálculo da probabilidade da união de três eventos é facilmente entendida com a ajuda do diagrama de Venn: $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + 2\,P(A \cap B \cap C)$.

Exemplo 2.4: Considere o experimento aleatório "jogar um dado e observar a face de cima". Sejam dois eventos: A "o número é maior do que 3", A = {4,5,6}; e B "o número é par" , B = {2,4,6}. O evento (A∩B)={4,6}. Assim, P(A) = 3/6, P(B) = 3/6 e P(A ∩ B) = 2/6.
Pela definição acima, P(A ∪ B) = 3/6 + 3/6 – 2/6 = 4/6.
Se usarmos o fato de que (A ∪ B)= {2,4,5,6} e a primeira definição de probabilidade, chegaremos ao mesmo resultado, pois

$$P(A \cup B) = \frac{\text{número de elementos do evento (A} \cup \text{B)}}{\text{total de elementos do espaço amostral}} = \frac{4}{6}$$

2.4. Probabilidade Condicional

A probabilidade de um evento A ocorrer dado que o evento B ocorreu é dada por:

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)}} \quad \text{, se P(B)> 0.} \qquad (1)$$

Exemplo 2.5: Considere o experimento aleatório "jogar um dado e observar a face de cima". Sejam dois eventos: F "o número é par", F = {2,4,6}; e B "o número é 2" , B={2}. Depois que o experimento é realizado, alguém informa que o evento F ocorreu, ou seja, a face é par. Dado que o evento F ocorreu, qual é a probabilidade de o evento B ocorrer? Ou seja, dado que o número é par, qual é a probabilidade de ser o número 2?

Usando definição (1), temos que

$$P(B \mid F) = \frac{P(F \cap B)}{P(F)} = \frac{P(\{2\})}{P(\{2,4,6\})} = \frac{1/6}{3/6} = \frac{1}{3}$$

Outra maneira de obter esse resultado é enxergar a informação da ocorrência do evento F como uma redução no espaço amostral inicial, E={1,2,3,4,5,6}. Se sabemos que o número é par, o novo espaço amostral para a ocorrência do evento B é E'={2,4,6}. Desse modo, podemos usar a definição clássica diretamente: o número de elementos favoráveis a B é 1 dentre os três resultados possíveis do espaço amostral reduzido E', sendo P(B | F) = 1/3.

2.5. Independência de Eventos

Quando o conhecimento sobre a ocorrência de um evento não altera a probabilidade de ocorrência de outro evento, esses eventos são ditos **independentes**. Ou seja, se a ocorrência de B não altera a probabilidade de ocorrência de A, os eventos A e B são independentes.

Dizemos que o evento A é independente do evento B se $P(A \mid B) = P(A)$.

Assim, pela definição (1), se os eventos A e B são independentes P(A ∩ B) = P(A) x P(B).

Exemplo 2.6: os eventos B e F do Exemplo 2.5 são exemplos de eventos **dependentes**, pois o conhecimento sobre a ocorrência do evento F aumentou a probabilidade do evento B, que antes era de 1/6. Ou seja P(B | F) ≠ P(B). Se B é dependente de F, também é dependente do complementar de F ($\bar{F}$), os números ímpares. Sabendo que o número é ímpar, a probabilidade de ser o número 2 é zero.

Exemplo 2.7: seja o seguinte experimento aleatório "observar o nascimento dos dois primeiros filhos de um casal e verificar o sexo das crianças" . O espaço amostral é E={ (F1 ∩ M2) , (F1 ∩ F2) , (M1 ∩ F2), (M1 ∩ M2) }, onde F1 denota menina no primeiro nascimento e M2 denota menino no segundo nascimento e assim por diante. Até onde se sabe, cada uma das quatro combinações que compõem o espaço amostral tem a mesma probabilidade de ocorrer, ou seja, ¼. Mas, será que a ocorrência de um menino no segundo nascimento (M2) é independente da ocorrência de uma menina no primeiro nascimento (F1)?

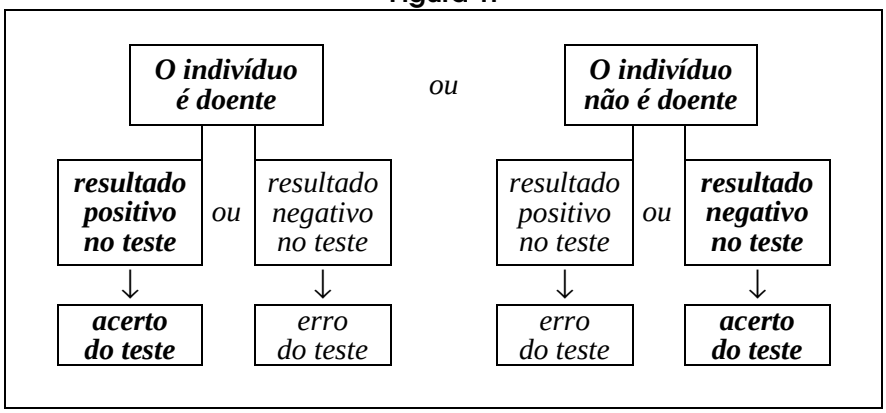
Vejamos: O evento F1 " menina no primeiro nascimento" é a união de dois dos eventos do espaço amostral, F1={(F1 ∩ M2) ∪ (F1 ∩ F2)} . Como não existe interseção entre esse dois eventos, P(F1) é dada pela soma da probabilidade dos dois eventos, isto é, 2/4. Usando o mesmo raciocínio, chegaremos à conclusão de que P(M2) também é igual a 2/4. A probabilidade da interseção de F1 e M2 é P(F1 ∩ M2) = ¼. Assim, verificamos que P(F1) x P(M2) = (2/4) x (2/4) = ¼ = P(F1 ∩ M2), então os eventos F1 e M2 são independentes.

3. Medidas da Qualidade do Teste Diagnóstico

Quando um teste diagnóstico é proposto para uma doença, ele deve passar por uma avaliação da sua qualidade de diagnosticar a doença ou a ausência dela.

A Figura 1 ilustra o que pode ocorrer quando o teste a ser avaliado é aplicado a amostras de duas populações, uma de pessoas sabidamente doentes e outra de não doentes. O resultado *positivo* no teste significa um diagnóstico de presença da doença e o resultado *negativo* significa um diagnóstico de ausência da doença.

Figura 1:



A definição da doença é feita por outro teste ou critério, chamado *padrão ouro*. O padrão ouro tem baíxissima possibilidade de erro, mas, em geral, é um teste caro, difícil de ser feito ou causador de desconforto para o paciente.

3.1. A Sensibilidade e a Especificidade de um Teste

As duas situações de acerto do teste diagnóstico mostradas na Figura 1 estão associadas às probabilidades chamadas de *sensibilidade* e *especificidade* do teste. Vamos adotar a seguinte notação para os eventos:

D: o indivíduo é doente,
$\overline{D}$: o indivíduo não é doente,
+: o indivíduo do teste é positivo,
- : o indivíduo do teste é negativo.

Sensibilidade (s): É a probabilidade do teste ter resultado positivo dado que o indivíduo está doente. Assim, na notação de probabilidade, temos

$$s = P(+ \mid D) = \frac{P(+ \cap D)}{P(D)} \tag{2}$$

que, na prática, é estimada pela proporção de resultados positivos do teste dentre os indivíduos sabidamente doentes.

Especificidade(e): É a probabilidade do teste teve resultado negativo dado que o indivíduo não está doente. Assim na notação de probabilidade, temos

$$e = P(- \mid \overline{D}) = \frac{P(- \cap \overline{D})}{P(\overline{D})} \tag{3}$$

que, na prática, é estimada pela proporção de resultados negativos do teste dentre os indivíduos sabidamente não doentes.

Assim, um teste muito sensível é útil para detectar a presença da doença em indivíduos doentes: se o indivíduo está doente, ele vai indicar isto com alta probabilidade. Por outro lado, um teste muito específico é um teste útil para excluir a presença da doença em indivíduos sadios: se o indivíduo não está doente, ele vai indicar isto com alta probabilidade.

Exemplo 1: O teste de Papanicolaou permite o diagnóstico precoce do câncer de colo de útero. Para avaliar a acurácia diagnóstica deste teste, realizou-se um estudo com 373 pacientes atendidas no Hospital das Clínicas de Botucatu (Pinho e Matos, 2002). O teste padrão-ouro utilizado em cada paciente para classificá-la em portadora ou não do câncer de colo de útero foi o exame histopatológico de uma amostra de tecido através de biópsia cervical. Os resultados são apresentados na Tabela 1.

Tabela 1: Resultados do teste Papanicolaou na detecção de câncer de colo de útero.

Resultado do teste Papanicolaou	Câncer de colo de útero		Total
	Doente	Não Doente	
Positivo	265	47	312
Negativo	11	50	61
Total	276	97	373

Baseados nesta amostra de 276 pacientes portadoras do câncer de colo útero, podemos estimar a sensibilidade do teste Papanicolaou:

$$s = P(+ \mid D) = \frac{\text{no. de pacientes doentes com resultado positivo}}{\text{no. de pacientes doentes}} = \frac{265}{276} = 0{,}96 \text{ ou } 96\%.$$

Do mesmo modo, através da amostra de 97 pacientes não portadoras do câncer de colo útero, podemos estimar a especificidade do teste Papanicolaou:

$$e = P(- \mid \overline{D}) = \frac{\text{no. de pacientes não doentes com resultado negativo}}{\text{no. de pacientes não doentes}} = \frac{50}{97} = 0{,}515 \text{ ou } 51{,}5\%.$$

Este teste tem alta sensibilidade mas especificidade muito baixa.

Muitas vezes, quando divulga-se a capacidade de acerto de um teste, é dado apenas o total de acertos dividido pelo total de indivíduos testados, sem separar os indivíduos sadios dos doentes. No exemplo anterior, esta porcentagem é de 84,5% ((265+50)/373).

Entretanto, esta capacidade de acerto *global* do teste, sem distinção entre acertos dentre os doentes (sensibilidade) e os acertos entre não doentes (especificidade), pode ser pouco informativa ou mesmo enganosa. Por exemplo, se um teste deu resultado positivo em 45 dos 50 indivíduos doentes, e resultado negativo em 15 dos 25 pacientes não doentes, a tal capacidade de acerto global do teste é de (45+15)/(50+25)= 0,80 (80%), que pode ser considerada alta. Mas note que, embora a capacidade de acerto do teste seja mesmo bastante alta entre os doentes (s = 0,90), não é alta entre os não doentes (e = 0,60). Informar apenas esta porcentagem global de acertos poderia levar à falsa conclusão de que o teste tem 80% de chance de acertar em qualquer caso (dado que o indivíduo é doente ou dado que o indivíduo é sadio), o que não é verdade.

3.2. A Influência do Ponto de Corte

Considere um exemplo genérico, onde a presença de uma doença é detectada por um teste através da ocorrência de valores altos de uma certa variável contínua medida nos pacientes, como, por exemplo, a medição de glicose no sangue para diagnosticar diabetes. A definição dos resultados *positivo* e *negativo* do teste é feita a partir da escolha de um *ponto de corte* na escala valores da variável: valores da medição acima do ponto de corte definem um resultado positivo e, abaixo, um resultado negativo no teste.

A Figura 2 mostra, para esta situação, a distribuição dos valores da variável entre os indivíduos sadios e entre os indivíduos doentes de uma população, considerando três pontos de corte.

Nesta figura, a sensibilidade do teste é a área na curva dos doentes que está acima do ponto de corte; a especificidade é a área na curva dos não doentes que está abaixo do ponto de corte.

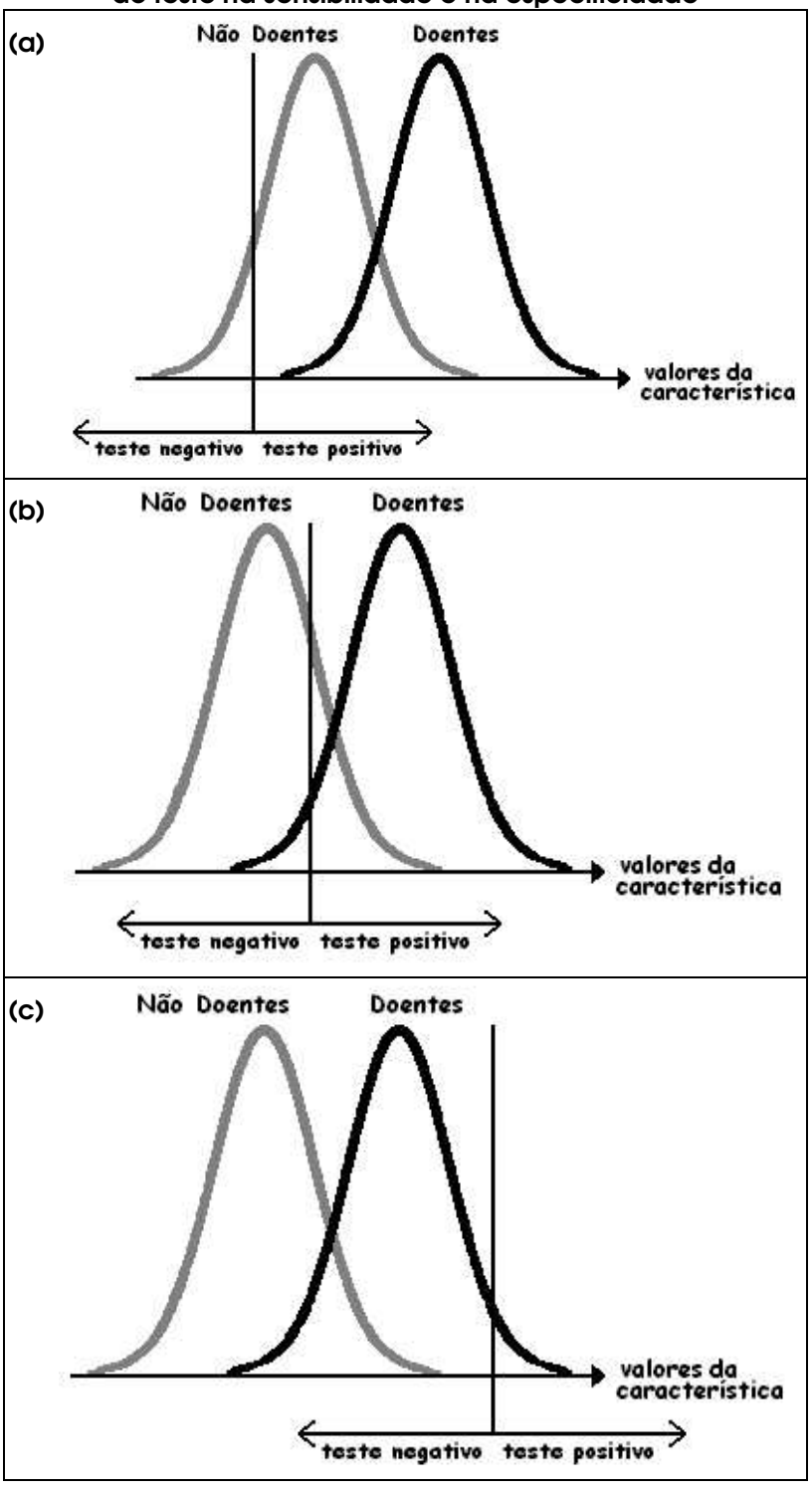
À medida que o ponto de corte aumenta, ou seja, passando da Figura 2(a) para a Figura 3(c), a sensibilidade diminui, pois estamos sendo mais rigorosos na definição de um doente ao exigir valores mais altos da variável para considerá-lo "positivo"; ao mesmo tempo, a especificidade aumenta.

Este exemplo mostra que, em relação à definição do ponto de corte, aumentar a sensibilidade de um teste implica necessariamente em reduzir sua especificidade, e vice-versa. Nos testes baseados na medição de variáveis quantitativas, a escolha do ponto de corte pode ser a ferramenta para tornar o teste mais sensível ou mais específico, conforme o objetivo do seu uso. Na seção 4.4, veremos como esta escolha influenciará na qualidade do diagnóstico baseado no teste.

□ **Exemplo 2:** Acredita-se que um valor alto para a razão da medida da cintura pela medida do quadril (RCQ) seja indicativo de hipertensão arterial. Pereira *at alli* (1999) fizeram um estudo com homens e mulheres de diversas faixas etárias. Foram consideradas hipertensas as pessoas com pressão sistólica igual ou maior que 140 mmHg ou pressão diastólica maior ou igual que 50 mmHg.

Vários pontos de corte do valor de RCQ foram considerados para definir o resultado positivo no teste (valor de RCQ acima do ponto de corte). O Quadro 1 reproduz os resultados mostrados no artigo para mulheres de 40 a 50 anos de idade, das quais 59 foram consideradas hipertensas e 326 não hipertensas.

Figura 2: A influência do ponto de corte de definição do resultado do teste na sensibilidade e na especificidade



Quadro 1: Reprodução de parte da Tabela 2 do artigo original de Pereira at alli (1999), referente aos resultados para mulheres de 40 a 50 anos de idade.

Teste RCQ é + se	Hipertensa e Teste RCQ +	Hipertensa e Teste RCQ -	Não hipertensa e Teste RCQ +	Não hipertensa e Teste RCQ -	Sensibi--lidade (%)	Especifi--cidade (%)
> 0,75	59	0	325	1	100	0,3
> 0,80	58	1	311	15	98,3	4,6
> 0,85	54	5	262	64	91,5	19,6
> 0,90	38	21	183	143	64,4	43,9
> 0,95	28	31	95	231	47,5	70,9
> 1,00	13	46	42	284	22	87,1
> 1,05	5	54	18	308	8,5	94,5
> 1,10	5	54	10	316	8,5	96,9
> 1,15	0	59	7	319	0	97,9

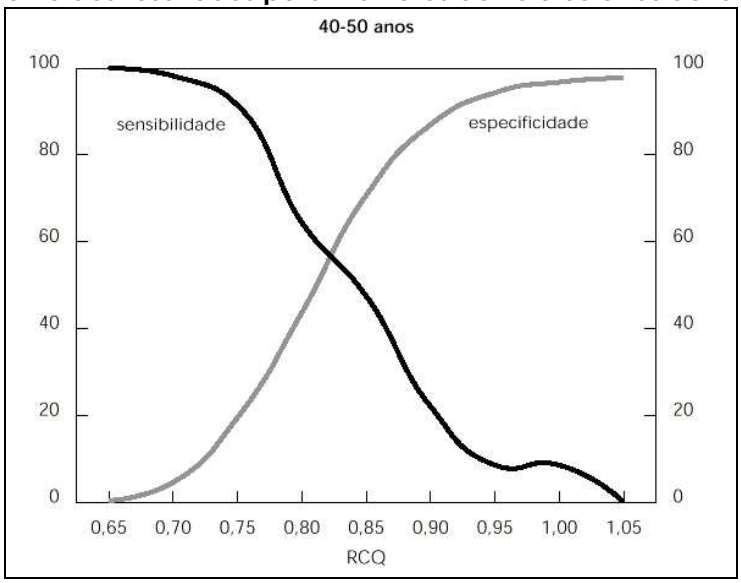
Os valores de sensibilidade e especificidade do Quadro 1 podem ser facilmente reproduzidos pelo leitor. Por exemplo, para ponto de corte RCQ > 0,85, temos:

$$s = P(+ \mid D) = \frac{\text{no. de hipertensas com resultado positivo}}{\text{no. de hipertensas}} = \frac{54}{54+5} = 0{,}915$$

$$e = P(- \mid \overline{D}) = \frac{\text{no. de não hipertensas com resultado negativo}}{\text{no. de hipertensas}} = \frac{64}{262+64} = 0{,}196\,.$$

A Figura 3 mostra as curvas de sensibilidade e especificidade em função dos diversos pontos de corte para o teste do RCQ. Verifica-se que, quanto maior o ponto de corte do RCQ, menor é a sensibilidade do teste. Isto ocorre porque se aumentamos o ponto de corte do RCQ, menos pessoas hipertensas serão consideradas positivas pelo teste, reduzindo a sensibilidade. Por outro lado, mais pessoas não hipertensas serão consideradas negativas pelo teste, aumentando sua especificidade, como podemos notar na Figura 3.

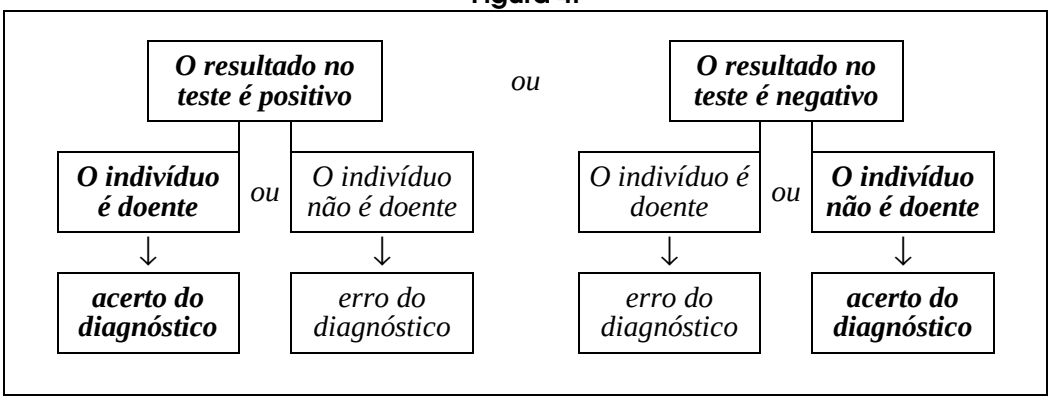
Figura 3: Reprodução de parte da Figura 2 do artigo original de Pereira at alli (1999), referente aos resultados para mulheres de 40 a 50 anos de idade.



4. Medidas da Qualidade do Diagnóstico Baseado em um Teste

Um vez que um teste clínico é adotado, devemos avaliar a qualidade do diagnóstico baseado nele. O usuário do teste não sabe se o indivíduo está doente ou não; a única informação que ele tem é o resultado do teste, positivo ou negativo. A Figura 4 ilustra o que pode ocorrer quando o teste a ser avaliado é aplicado a um indivíduo.

Figura 4:



4.1. Os Valores de Predição Positiva e Negativa

Às duas situações de acerto do diagnóstico baseado no teste mostradas na Figura 4 estão associadas probabilidades chamadas de *valor de predição positiva* e *valor de predição negativa* do diagnóstico. Vejamos suas definições a seguir.

Valor de Predição Positiva (VPP): É a probabilidade do indivíduo estar doente dado que o seu teste teve resultado positivo. Assim:

$$VPP = P(D \mid +) = \frac{P(D \cap +)}{P(+)} \tag{4}$$

Valor de Predição Negativa (VPN): É a probabilidade do indivíduo não estar doente dado que o seu teste teve resultado negativo. Assim:

$$VPN = P(\overline{D} \mid -) = \frac{P(\overline{D} \cap -)}{P(-)} \tag{5}$$

Desse modo, o VPP e o VPN são, respectivamente, a probabilidade de acerto do diagnóstico positivo e a probabilidade de acerto do diagnóstico negativo do teste. As probabilidades dos eventos complementares deste acertos são chamadas Proporção de Falsos Positivos e Proporção de Falsos Negativos, definidas a seguir.

Proporção de Falsos Positivos (PFP): É a probabilidade do paciente não estar doente dado que o teste teve resultado positivo. Assim:

$$PFP = P(\overline{D} \mid +) = \frac{P(\overline{D} \cap +)}{P(+)} \quad \text{ou seja,} \quad PFP = 1 - VPP$$

Proporção de Falsos Negativos (PFN): É a probabilidade do paciente estar doente dado que o teste teve resultado negativo. Assim:

$$VPN = P(D \mid -) = \frac{P(D \cap -)}{P(-)} \quad \text{ou seja,} \quad PFN = 1 - VPN$$

4.2. Estimando os Valores de Predição

A proporção de indivíduos portadores da doença na população é chamada de **prevalência** da doença na população, denotada por **p**. A prevalência p também pode ser interpretada como a probabilidade de um indivíduo, selecionado aleatoriamente desta população, ser doente, ou seja P(D) = p. Desse modo, a probabilidade do indivíduo não ser doente é dada por P($\overline{D}$)= 1 - p.

O VPP e o VPN do diagnóstico baseado no teste depende: da sensibilidade e da especificidade do teste, e da prevalência da doença na população. Deste modo, conhecendo *s*, *e* e *p*, podemos calcular o VPP e o VPN através das seguintes equações[2]:

$$\boxed{VPP = \frac{sp}{sp+(1-e)(1-p)}} \qquad (6)$$

e

$$\boxed{VPN = \frac{e(1-p)}{e(1-p)+(1-s)p}} \qquad (7)$$

Exemplo 1 (continuação): Vamos voltar ao teste Papanicolaou para diagnóstico do câncer de colo de útero apresentado anteriormente. Considerando uma população com uma prevalência de câncer de útero de 5% (p=0,05), os valores de predição positiva e negativa seriam calculados da seguinte forma:

$$VPP = \frac{0{,}96\times 0{,}05}{0{,}96\times 0{,}05+(1-0{,}515)(1-0{,}05)} = 0{,}094 \qquad \text{e} \qquad VPN = \frac{0{,}515(1-0{,}05)}{0{,}515(1-0{,}05)+(1-0{,}96)0{,}05} = 0{,}996$$

Assim, se aplicado a uma população com uma prevalência de câncer de útero de 5%, uma paciente com resultado negativo no teste Papanicolaou tem uma probabilidade muito alta (99,6%) de não ser portadora do câncer de colo de útero. No entanto, uma paciente com resultado positivo tem uma probabilidade muito pequena de ser realmente portadora da doença (9,4%). Desse modo, o resultado positivo do teste Papanicolaou é pouco conclusivo, mas o resultado negativo é muito confiável.

Vamos considerar o VPP e o VPN do teste Papanicolaou aplicado a uma população com prevalência maior de câncer do colo do útero, digamos, p = 0,20:

$$VPP = \frac{0{,}96\times 0{,}20}{0{,}96\times 0{,}20+(1-0{,}515)(1-0{,}20)} = 0{,}664 \qquad \text{e} \qquad VPN = \frac{0{,}515(1-0{,}20)}{0{,}515(1-0{,}20)+(1-0{,}96)0{,}20} = 0{,}981.$$

Note que há um aumento significativo no VPP e uma queda pequena no VPN. Na próxima seção discutiremos o efeito da prevalência, da sensibilidade e da especificidade nos valores de predição positiva e negativa.

4.3. O Efeito da Prevalência nos Valores de Predição

Através das equações (6) e (7), reproduzidas abaixo, podemos ver que, para um determinado teste (ou seja, para valores fixos de sensibilidade e especificidade), quando maior a prevalência da doença maior será o VPP e menor será o VPN:

$$VPP = \frac{sp}{sp+(1-e)(1-p)} \qquad \text{e} \qquad VPN = \frac{e(1-p)}{e(1-p)+(1-s)p}.$$

[2] Você pode derivar estas equações a partir das definições (4) e (5), escrevendo-as em função das definições (2) e (3).

Para "visualizar" este efeito da prevalência nos valores de predição, observe a Figura 5. Neste gráfico, para avaliar apenas o efeito da prevalência no VPP e no VPN fixamos os valor da sensibilidade em 0,8 e da especificidade em 0,7:

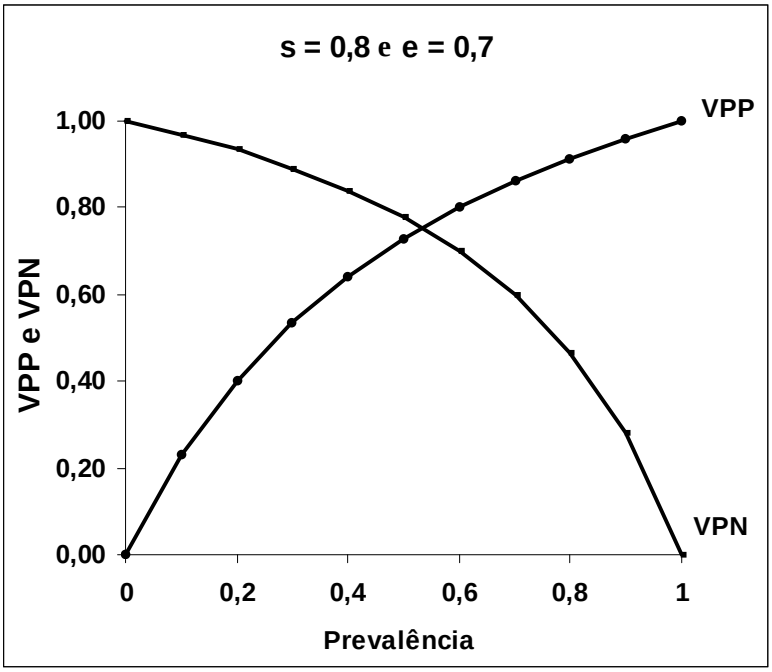
$$VPP = \frac{0{,}8p}{0{,}8p + 0{,}3(1-p)} \qquad e \qquad VPN = \frac{0{,}7(1-p)}{0{,}7(1-p) + 0{,}2p}.$$

Agora, nas equações acima, VPP e VPN dependem apenas da prevalência: para cada valor de p termos um valor diferente de VPP e VPN. Por exemplo, fazendo p = 0,4, temos:

$$VPP = \frac{0{,}8(0{,}4)}{0{,}8(0{,}4) + 0{,}3(1-0{,}4)} = 0{,}64 \qquad e \qquad VPN = \frac{0{,}7(1-0{,}4)}{0{,}7(1-0{,}4) + 0{,}2(0{,}4)} = 0{,}84.$$

Na Figura 4, mostramos os valores de VPP e VPN fazendo a prevalência variar entre 0 e pode-se ver claramente que quanto maior a prevalência da doença, maior será o VPP e menor será o VPN do disgnóstico.

Figura 5: Efeito da Prevalência no VPP e no VPN.



Exemplo 3: Atualmente, a presença do HIV-AIDS é detectada rotineiramente pelo teste ELISA que tem sensibilidade de 95,0% e especificidade de 99,8% (Laboratório ABBOTT). Em novembro de 2002, o Laboratório OraSure lançou o teste OraQuick® (o resultado fica pronto em 20 minutos) com sensibilidade de 99,6% e especificidade de 100,0%.

A Tabela 2 mostra os valores de predição positiva e negativa do diagnóstico do HIV-AIDS baseado nestes testes para crescentes níveis de prevalência da infecção.

Como podemos notar, especialmente para o teste ELISA, à medida que a prevalência cresce, ocorre aumento no VPP e diminuição do VPN. Isso significa que o diagnóstico baseado no teste ELISA tem uma grande proporção de falsos positivos em populações com baixa proporção de infectados pelo HIV-AIDS, de modo que um resultado positivo deve ser investigada com repetições do teste ou com outro teste. Por outro lado, o resultado negativo tem um valor de predição muito alto mesmo em grupos de altíssima prevalência. Esta é uma característica dos testes de triagem, onde os resultados negativos devem ser altamente confiáveis, enquanto os casos positivos são triados para novos testes.

No caso do teste OraQuick®, é possível observar somente um pequeno decréscimo no VPN para grupos de altíssimo risco. O VPP será sempre o valor máximo de 100% devido ao fato de que

o teste identifica todas as pessoas não infectadas (especificidade igual a 100%). Esta influência da sensibilidade e da especificidade nos valores de predição é assunto da próxima seção.

Tabela 2: Valores de Predição Positiva e Negativa no diagnostico do HIV-AIDS do teste Elisa e do teste OraQuick®.

Prevalência	ELISA (ABBOTT)		OraQuick®	
	VPP (%)	VPN (%)	VPP (%)	VPN (%)
1 em 1 milhão	0,05	100	100	100
1 em 500 mil	0,09	100	100	100
1 em 100 mil	0,47	100	100	100
1 em 50 mil	0,94	100	100	100
1 em 10 mil	4,54	100	100	100
1 em 5 mil	8,68	100	100	100
1 em mil	32,23	99,99	100	100
1 em 500	48,77	99,99	100	100
1 em 100	82,75	99,95	100	100
1 em 50	90,65	99,90	100	99,99
1 em 10	98,14	99,45	100	99,96

4.4. Estudando o Efeito da Sensibilidade e da Especificidade nos Valores de Predição

Através das equações (6) e (7), reproduzidas abaixo, podemos ver que, para prevalência fixa, quanto maior o valor da sensibilidade e/ou da especificidade, maior será o VPP e o VPN:

$$VPP = \frac{sp}{sp+(1-e)(1-p)} \quad \text{e} \quad VPN = \frac{e(1-p)}{e(1-p)+(1-s)p}.$$

Entretanto, vamos ver que o VPP é mais afetado pela especificidade do que pela sensibilidade e que o VPN, ao contrário, é mais afetado pela sensibilidade.

A Figura 6(a) mostra o efeito da especificidade no VPP, para prevalência fixa em 5% e três valores fixos de sensibilidade, correspondentes a cada uma das curvas. Note que, em cada curva (sensibilidade fixa), quando a especificidade sobe de 0,75 para 0,95, o VPP é aumentado em 0,30, enquanto que, para um valor fixo de especificidade, passando de uma curva para outra (aumento a sensibilidade de 0,75 para 0,95), ao VPP é acrescido de apenas 0,08, aproximadamente. O gráfico mostra que o mesmo aumento na especificidade ou na sensibilidade produz efeitos diferentes no VPP. Assim, se desejamos aumentar o valor de predição positiva de um diagnóstico (reduzir a probabilidade de falsos positivos), devemos investir no aumento da especificidade do teste.

O efeito inverso é mostrado na Figura 6(b): o mesmo aumento na sensibilidade ou na especificidade produz efeitos diferentes no VPN, efeito este muito maior no caso de aumento na sensibilidade. Desse modo, se desejamos aumentar o valor de predição negativa de um diagnóstico (reduzir a probabilidade de falsos negativos), devemos investir no aumento da sensibilidade do teste.

Juntamente com a análise gráfica feita anteriormente, podemos ver como a especificidade tem mais efeito no VPP do que no VPN considerando um caso extremo de um teste com especicifidade em seu valor máximo. Se $e=P(-|ND)=1$, o resultado do teste é negativo para todos os não-doentes. Como não existem não-doentes com teste positivo, todos os positivos são de doentes: $P(D|+) = 1 = VPP$, ou seja, conseguimos fazer o VPP atingir seu valor máximo. Entretanto, não conseguimos este efeito no $VPN=P(ND|-)$, pois, mesmo que a especificidade seja máxima, nem todos os negativos seriam de não-doentes.

O mesmo raciocínio pode ser usado para ver que a sensibilidade influencia mais o VPN que no VPP. Se $s = P(+|D) = 1$, o teste dá positivo para todos os doentes; então não há doentes com teste negativo, ou seja, todos os negativos são de não-doentes: $P(ND|-) = 1= VPN$. Mas não conseguimos que $VPP = P(D|+)$ seja igual a 1, pois nem todos os positivos seriam de doentes.

Figura 6(a): Efeito da especificidade no Valor de Predição Positiva

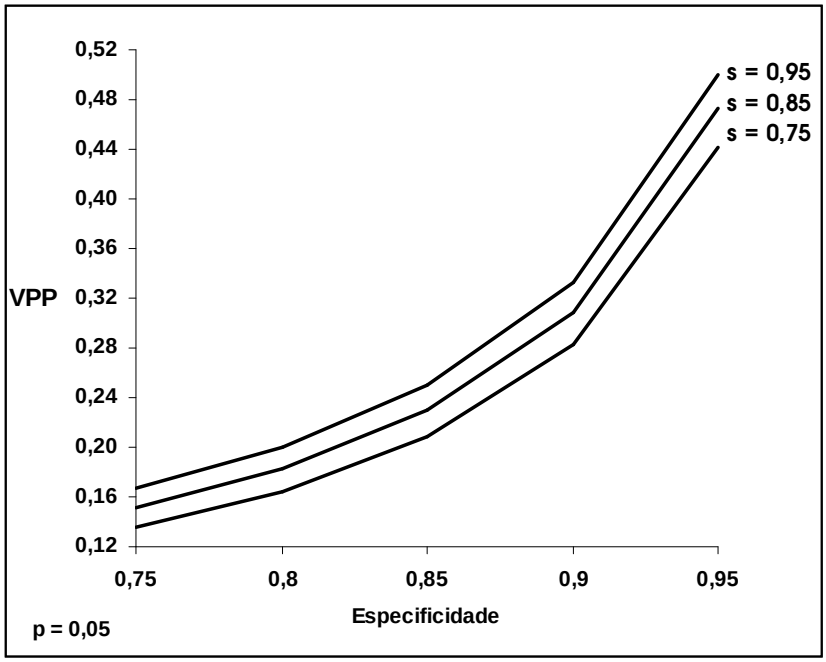
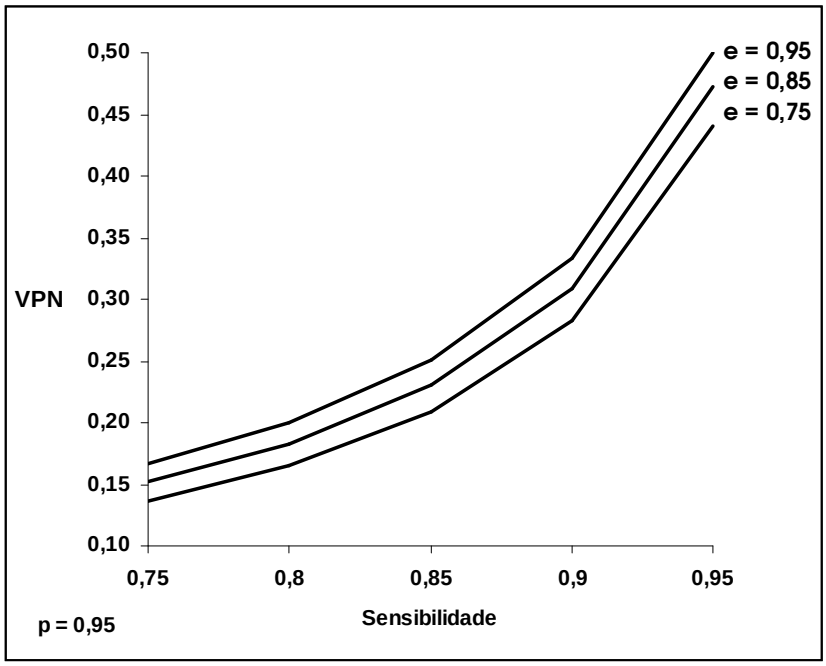


Figura 6(b): Efeito da sensibilidade no Valor de Predição Negativa



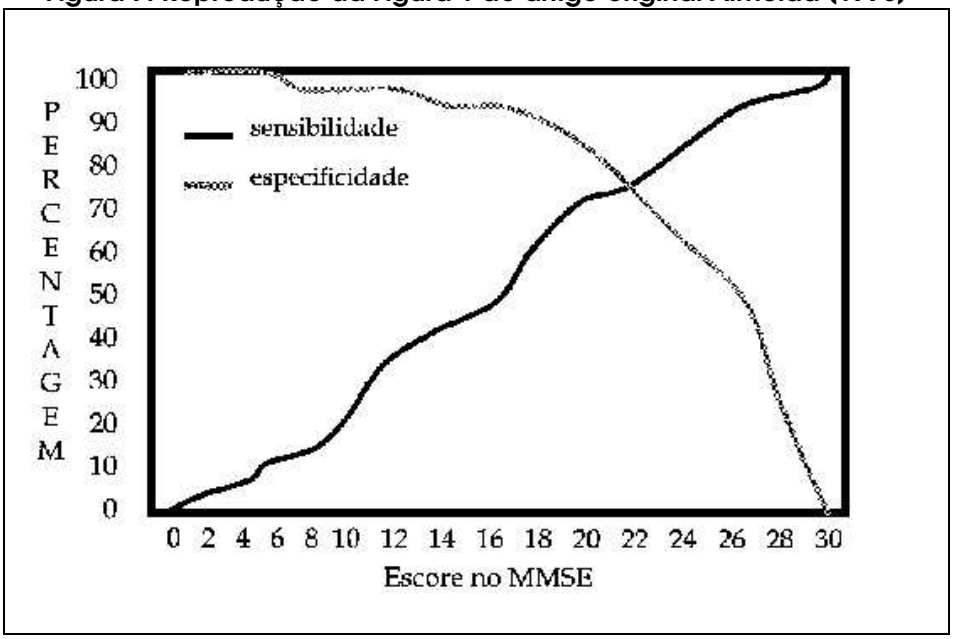
Exemplo 4: O Mini Exame do Estado Mental (MEEM) é um teste usado para diagnosticar demência. O MEEM é composto por diversas questões com o objetivo de avaliar funções cognitivas específicas do paciente. Seu escore é um número inteiro, podendo variar de 0 a 30 pontos: quanto menor o escore, maior o indício de demência. A escala é simples de usar e pode ser facilmente administrada em 5 a 10 minutos, inclusive por profissionais não médicos.

Com o objetivo de investigar o melhor ponto de corte na escala do MEEM para classificar o paciente como portador ou não de demência, Almeida (1998) realizou um estudo com 211 pacientes idosos, dos quais 70 receberam diagnóstico de demência de acordo com as diretrizes da CID-10 (o teste padrão-ouro neste caso)[3].

[3] Classificação Internacional de Doenças, que é um catálogo com a classificação de doenças, usado como referência em todo o mundo.

Os 30 possíveis valores do MEEM foram avaliados quanto à sua sensibilidade e especificidade para o diagnóstico de demência, gerando a Figura 7.

Figura 7: Reprodução da Figura 1 do artigo original Almeida (1998)



O ponto de corte tradicional ("positivo" se menor ou igual a 23 pontos, o MEEM-23) revelou sensibilidade de 84,3% e especificidade de 60,3%. Segundo o artigo, a prevalência de demência entre idosos de 60 a 80 anos é de 5% Assim, aplicando o MEEM-23 a um paciente com idade entre 60 e 80 anos, o VPP será de 10% e o VPN de 98,6%. Portanto, o MEEM-23 com ponto de corte no escore 23 é útil para descartar a demência, ou seja, é satisfatório como um teste de triagem (que é o seu objetivo, já que é mais rápido e mais simples que a classificação baseada na CID-10).

Entre idosos com mais de 80 anos, a prevalência de demência é de 20%. Desse modo, aplicando o MEEM-23 a um paciente nesta faixa etária, o VPP sobe para 34,7% e o VPN sofre uma queda pequena, indo para 93,9%. Nesta situação, o MEEM-23 continua sendo um teste útil apenas para triagem. Entretanto, se um VPN de 93,9% não for considerado suficientemente alto, este valor pode ser aumentado se a sensibilidade do MEEM for aumentada, o que pode ser conseguido abaixando-se o ponto de corte.

5. Outros Exemplos Interessantes

Exemplo 5: A associação entre a presença da bactéria *Helicobacter pylori* (HP) no trato estomacal e a existência de úlcera péptica foi estabelecida na década de 80 e, desde então, a questão de como detectar a HP em exames de rotina ganhou grande importância. O procedimento ideal para detecção da HP é realizar uma endoscopia digestiva, que é um exame caro e causador de muito desconforto para o paciente.

Outro procedimento, chamado UBT (*Urea Breath Test*), também pode ser utilizado para detectar a HP. Para se submeter ao UBT, o paciente ingere uma solução com um isótopo de carbono marcado. Se o paciente possui a HP em seu trato estomacal, espera-se que o isótopo de carbono seja quebrado em amônia e CO_2 marcado, que é absorvido no sangue e expelido na respiração. Amostras de CO_2 expelido são coletadas e analisadas para a presença do isótopo. Dois isótopos de carbono podem ser usados: ^{14}C, um isótopo radioativo, e ^{13}C, não radioativo.

Gomes e colegas (2002) realizaram um estudo para avaliar a qualidade do UBT feito com o ^{14}C, que denominaram ^{14}C-UBT. O estudo envolveu 137 pacientes do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (USP), dos quais 115 eram portadores da HP. O teste padrão-ouro utilizado foi a endoscopia digestiva.

O resultado do ^{14}C-UBT foi medido pela radioatividade do ar expelido pelo paciente e expresso em contagens por minuto (cpm). Foram feitas duas coletas de ar: uma 15 minutos depois

da ingestão da solução com o ^{14}C e outra 30 minutos depois da ingestão. Três pontos de corte para a medição da radioatividade foram estudados: 1000 cpm, 1500 cpm e 2000 cpm. O resultado é considerado positivo se o valor da medição é maior do que o ponto de corte utilizado.

O Quadro 2 reproduz parte dos resultados apresentados no artigo original.

Quadro 2: Reprodução de parte da Tabela 2 do artigo original de Gomes e colegas (2002)

Tempo após a ingestão	1000 cpm		1500 cpm		2000 cpm	
	Sensibilidade	Especificidade	Sensibilidade	Especificidade	Sensibilidade	Especificidade
15 minutos	99%	95%	97%	95%	94%	100%
30 minutos	99%	95%	96%	100%	90%	100%

Podemos notar que, em ambos os tempos de coleta, a sensibilidade do teste cai com o aumento do ponto de corte, ocorrendo o inverso com a especificidade. Como já sabemos, isto ocorre porque, com o aumento do ponto de corte, diminui o número de pacientes com resultado positivos para o teste e aumenta o número de negativos, causando a queda da sensibilidade e o aumento da especificidade.

Exemplo 6: Estudos epidemiológicos trazendo informações ocupacionais obtidas em entrevistas com familiares de trabalhadores são muito freqüentes. Entretanto, há poucos trabalhos avaliando a qualidade destas informações.

Cordeiro (2000) realizou um estudo no qual 2460 pessoas foram questionadas sob sua história ocupacional, e a mesma informação sobre esta pessoa foi pedida a seus famillliares. Nesta situação, o "teste" é a informação dada pelo familiar. A informação fornecida pela própria pessoa é o padrão-ouro.

O Quadro 3 reproduz a Tabela 2 do artigo original, que mostra os valores de sensibilidade e especificidade da informação fornecida pelo familiar para vários subgrupos ocupacionais.

Quadro 3: Reprodução da Tabela 2 do artigo de Cordeiro (2000)

Tabela 2 – Sensibilidade e especificidade das informações subgrupo de ocupação principal. Botucatu, SP, Brasil. 1998.

Subgrupo ocupacional*	Sensibilidade (IC95%) %	Especificidade (IC95%) %
Administrativos	79,3 (70,5-88,0)	99,7 (99,4-99,9)
Cientistas	92,3 (86,8-97,8)	99,9 (99,7-100,0)
Técnicos	88,9 (82,4-95,4)	99,5 (99,2-99,8)
Lavradores	87,4 (84,6-90,2)	96,9 (96,0-97,7)
Comerciários	82,4 (76,8-88,0)	98,8 (98,4-99,3)
Motoristas	87,3 (81,0-93,5)	99,6 (99,3-99,8)
Prestação de serviços	84,5 (80,3-88,7)	97,9 (97,3-98,6)
Escritório	81,9 (75,2-88,6)	99,5 (99,1-99,8)
Metalúrgicos	92,6 (88,2-97,0)	99,1 (98,6-99,5)
Construção civil	92,4 (87,6-97,2)	99,2 (98,8-99,6)
Têxteis	88,2 (72,9-100,0)	99,9 (99,7-100,0)
Vestuário	91,1 (85,2-97,0)	99,3 (99,0-99,7)
Madeira	88,2 (77,4-99,1)	99,8 (99,6-100,0)
Braçais	77,5 (64,6-90,4)	99,7 (99,4-99,9)
Sem ocupação	98,9 (97,3-100,0)	97,8 (97,2-98,5)

*Segundo Rumel[13] (1985).

Um resultado interessante é que o mais alto valor de sensibilidade da informação do familiar ocorre quando a pessoa se classifica como "sem ocupação". Isso significa que, dado que a pessoa nãotem ocupação, o familiar consegue identificar esta situação com maior facilidade. Por outro lado, a menor sensibilidade ocorre quando a pessoa se classifica como "trabalhador braçal", talvez pelo fato desta categoria englobar outras, como lavrador, construção civil, etc.

Referências Bibliográficas

Almeida, O.P. (1998) ´Mini Exame do Estado Mental e Diagnóstico de Demência no Brasil´. *Arquivos de Neuropsiquiatria*, v. 56, n. 3-B, p. 605-612.

Cordeiro, R. (2000) ´Validade de Histórias Ocupacionais de Informantes Secundários´. Revista de Saúde Pública, v. 34, n. 5, p. 522-528.

Gomes, A.T.B., Coelho, L.K., Secaf, M., Modena, J.L.P., Troncon, L.E.A. e Oliveira, R.B. (2002) ´Accuracy of ^{14}C-urea test for the diagonis of Helicobacter pylori' . *São Paulo Medical Journal – Revista Paulista de Medicina*, volume 120, no. 3, página 68-71.

Pereira, R.A., Sichieri, R. e Marins, V.M.R. (1999) ´Razão cintura/quadril como preditor de hipertensão arterial´. *Cadernos de Saúde Pública*, v. 15, n. 2, p. 333-344.

Pinho, A.A. e Matos, M.C.F.I (2002) ´Validade da citologia cervicovaginal na detecção de lesões pré-neoplásicas e neoplásicas de colo de útero´. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, v. 38, n. 3, p. 225-231.