SALUTIS SCIENTIA REVISTA DE CIÊNCIAS DA SAÚDE DA ESSCVP

VOL.2 – MARÇO 2010

Artigo de Opinião

Devemos acreditar em análises clínicas?(1)

Can we trust diagnostic tests?

Dinis Pestana [1,2*]

[1] Centro de Estatística e Aplicações da Universidade de Lisboa;
[2] Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa.

A Teoria da Probabilidade foi criada para domesticar a incerteza, transformando-a de inimiga em aliada; sendo a linguagem da Estatística, é um instrumento incontornável no tratamento da informação, em todas as ciências indutivas. Mas a Probabilidade é, em si mesma, um poderoso auxiliar na interpretação dos factos. Discute-se o teorema de Bayes, um resultado central que está no cerne da controvérsia sobre causalidade/ associação estatística, começando por uma apresentação concreta associada à interpretação de análises clínicas. Refere-se ainda o papel da meta-análise na obtenção de curvas ROC (*Receiver Operating Characteristics*), e como são usáveis na obtenção dos pontos de corte para distinguir positivos de negativos (verdadeiros ou falsos!) nas análises clínicas.

Probability Theory has a special role in taming uncertainty, and the mastery of risk is the inprint of modern science. As the language of Statistics, Probability has a major share in the methodology of experimental sciences, and is a cornerstone of indutive logic. On the other hand, probability is by itself a powerful tool in the interpretation of facts. We discuss Bayes theorem, a central result at the core of the statistical association/ causation debate. After an almost empirical presentation in the context of interpretation of results of a clinical test, we also discuss the obtention of ROC (Receiver Operating Characteristics) curves with the help of meta-analysis, and how they are used to define cut-points in a clinical analysis.

PALAVRAS-CHAVE: *Teorema de Bayes; causalidade; inferência; análises clínicas; sensibilidade; especificidade; valor preditivo positivo.*

KEY WORDS: *Bayes theorem; causality; inference; clinical tests; sensibility; specificity; positive predictive value.*

* **Correspondência:** Dinis Pestana. **Email:** dinis.pestana@fc.ul.pt

INTRODUÇÃO

Um dos preconceitos do nosso tempo é o do primado da informação ("*informação é poder*"), e o decorrente protagonismo das tecnologias da informação. No entanto, informação não é conhecimento, e informação mal compreendida pode ser mais prejudicial do que benéfica — recomenda-se por exemplo a leitura divertida e potencialmente proveitosa de *Ciência da Treta* (Goldacre, 2008), em que a discussão sobre confundimento, publicidade enganosa, Estatística mal feita, sonegação de informação, ou pior, substituição de informação correcta por informação irrelevante apresentada como fundamental, é feita de forma inquestionável. Disponibilizar informação em excesso, aliás, é uma das mais sofisticadas e perversas formas de não dar informação (Edgar Allan Poe escreveu um conto admirável sobre a melhor forma de ocultar uma carta importante e confidencial: entre muitas outras cartas, evidentemente!).

Um exemplo de como informação, em si mesma, tem um valor questionável para tomar decisões racionalmente: qualquer mulher fica com certeza em pânico se lhe disserem que tem cancro de mama, com a informação corrente de que é uma das principais causas de morte na nossa sociedade. Com a informação do número de óbitos por dia decorrentes dessa situação clínica, acede quase certamente a fazer uma mastectomia. No entanto, a maior parte dos cancros de mama são do tipo carcinoma *in situ*, que não evolui para fora dos canais deferentes de aleitação, e a manipulação de probabilidade condicional permite-nos calcular que o número de casos que é preciso tratar (*cases needed to treat*), com esse tipo de cirurgia mutilante, para salvar **uma** vida, é de **alguns milhares**! (Gigerenzer, 2002).

Mas como não queremos centrar a discussão num exemplo assim tão deprimente, vamos antes, na secção 1, discutir uma doença rara, a porfíria aguda intermitente, e o uso de uma análise como meio auxiliar de diagnóstico. A reflexão sobre os diversos casos apresentados, e em particular do caso III, torna bem patente que a informação (resultado positivo da análise) pode ser quase irrelevante. O que é importante, para que a análise tenha um valor preditivo digno de nota, é que seja interpretada por quem sabe extrair conhecimento da informação, e não a informação em si de que o resultado foi positivo. Espera-se que os leitores mais informados ampliem o que fica escrito, nomeadamente fazendo o paralelo entre sensibilidade e especificidade das análise clínicas com a probabilidade de erro de primeira espécie e a probabilidade de erro de segunda espécie dos testes de hipóteses.

Na secção 2 a apresentação é reformulada em termos do Teorema de Bayes, e chama-se a atenção para este extraordinário resultado, fundamental em tantas áreas de investigação. Desde que Pearson (1892, reedição recente: 2009) publicou o seu notável *The Grammar of Science* que o paradigma das ciências experimentais hesita entre investigar associação estatística ou causalidade; a obra notável de Fisher (veja-se a reedição conjunta de 1990), e nomeadamente os seus livros que estabeleceram as bases do planeamento experimental, ensinaram-nos que dados se deve analisar quando em vez de estudos meramente observacionais se pretende fazer estudos experimentais. Mas o debate sobre a causalidade prossegue, e é cada vez mais um ponto fundamental da Filosofia da Ciência e metodologia da investigação experimental, como único garante de indução fiável.

Finalmente na Secção 3, mais especializada e árida, procuramos satisfazer os leitores mais inquisitivos, e que porventura tenham ficado insatisfeitos com a aparente arbitrariedade da forma como a análise apresentada na Secção 1 decide quem é positivo e quem é negativo.

1. Valor preditivo da análise do teor sérico de diaminase de porfibilogénio

A porfíria aguda intermitente é uma doença transmitida geneticamente (carácter dominante), felizmente muito rara — estima-se que a prevalência na população europeia é $p = \frac{1}{10\,000}$.

O diagnóstico não é fácil, pois as manifestações clínicas são muito diversas; pode, durante anos, não ter qualquer manifestação, mas de repente

desencadear um comportamento socialmente pouco comum (muito bem explorado no filme *A Loucura do Rei George*) seguido de recuperação; mas também pode, logo na sua primeira manifestação, induzir coma profundo e irrecuperável, como no relato (*Paula*) que Isabel Allende fez da morte da sua filha. A porfíria aguda está associada a produção deficitária de diaminase de porfibilogénio, pelo que a determinação de níveis séricos desta enzima serve, naturalmente, como meio auxiliar de diagnóstico.

As análises clínicas são meios auxiliares de diagnóstico a que hoje se recorre rotineiramente. Por exemplo, para diagnosticar a porfíria aguda intermitente procedemos à medição do nível de diaminase de porfibilinogénio. Se este for inferior a 99 u/mm^3 o indivíduo é *positivo* no que refere a porfíria aguda intermitente, se for superior a 99 é considerado *negativo*. A análise referida **não** é um diagnóstico.

Era bom que as análises clínicas fossem infalíveis. Infelizmente qualquer "população" é em geral uma mistura de subpopulações, em que as fronteiras são ainda mais ténues do que as fronteiras geográficas: um "habitante" da área dos positivos pode não ter a doença — é um *positivo falso*, *PF* (e denotamos *PV* os positivos verdadeiros) — e um "habitante" da área dos negativos pode ter a doença — é um *negativo falso*, *NF*. Denotamos *NV* os negativos verdadeiros:

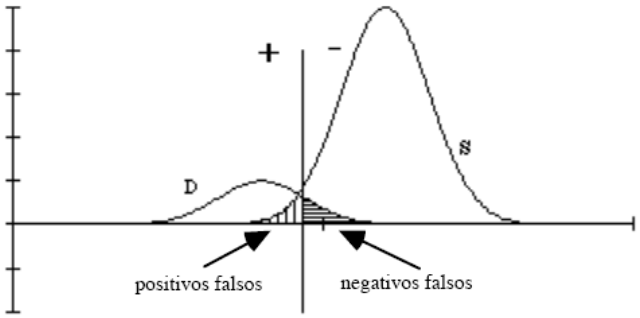
	doente	são	total
positivo	PV	PF	P
negativo	NF	NV	N
total	D	S	n

Claro que gostaríamos de só ter positivos verdadeiros, isto é que a sensibilidade da análise, a fracção de positivos entre os doentes *D*, $s = \frac{PV}{PV+NF} = \mathrm{P}(+|D)$, fosse 100%. Gostaríamos de não ter positivos falsos, isto é que a análise tivesse especificidade, a fracção de negativos entre os sãos $S = \overline{D}$, $e = \frac{NV}{PF+NV} = \mathrm{P}(-|\overline{D}) = 100\%$, só apontasse como positivos os efectivamente doentes.

A prevalência da doença, acima referida, é avaliada pela fracção $p = \frac{D}{n}$.

Mas, tal como um habitante de Espanha pode ser português, e há espanhóis a viver em Portugal, a situação não é em geral tão clara. A Figura 1 ilustra a situação (exagerando muito a curva correspondente aos doentes, que deveria ser quase invisível dada a baixa prevalência da doença).

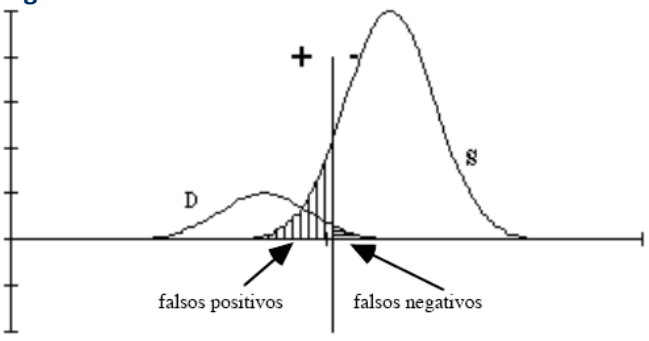
Figura 1 - Diagnóstico e situação real.



No caso da análise para detecção da porfíria, a fasquia foi colocada no nível 99 da enzima atrás referido; considerando positivos os casos em que o nível é inferior a 99 e negativos aqueles em que é superior, a sensibilidade da análise é 82% e a sua especificidade é 96.3%.

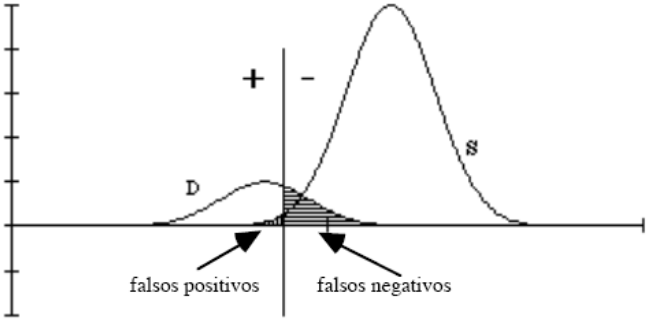
Claro que se deslocarmos a fasquia para a direita (Figura 2) a sensibilidade da análise aumenta, e no limite conseguiremos que todos ou quase todos os doentes sejam detectados como positivos — mas o preço a pagar é que simultaneamente começa a haver muitos positivos falsos.

Figura 2 - Maior sensibilidade.



Por outro lado, deslocando a fasquia para a esquerda (Figura 3) a especificidade da análise aumenta, e cada vez será menor a probabilidade de se diagnosticar erradamente a doença a um indivíduo são — mas então haverá uma probabilidade maior de um doente escapar à detecção através desta análise.

Figura 3 - Maior especificidade.



Encontraremos constantemente esta situação de compromisso no que respeita sensibilidade e especificidade. Parece que não se pode ter o melhor de dois mundos, e por isso temos que decidir qual dos riscos queremos controlar, ou chegamos a um compromisso sobre o equilíbrio desses riscos, como descrito na Secção 3.

Poderemos, nestas circunstâncias, confiar no resultado da análise? Por outras palavras, qual é o valor preditivo da análise? Convém, evidentemente, distinguir o valor preditivo positivo $\mathrm{P}(D\,|+)$ e o valor preditivo negativo $\mathrm{P}(\overline{D}\,|-)$.

Analisemos três casos possíveis em que se recorre a esta análise:

Caso I: Um médico suspeita que um dos seus pacientes, o Arnesto, sofre de porfíria, devido ao quadro clínico que observa. Estudos anteriores levam-nos a admitir que 30% dos indivíduos que apresentam o quadro clínico observado têm porfíria. Por isso manda o Arnesto fazer uma análise ao teor sérico de diaminase de porfibilogénio, que vem a revelar-se positiva. Qual é a probabilidade de o Arnesto ter de facto porfíria?

Em Estatística não se pode inferir para um indivíduo particular. Pode, por outro lado, avaliar-se a incerteza para um colectivo de indivíduos, e daí dar indicações sobre o que podemos esperar para um indivíduo seleccionado ao acaso nesse colectivo — não é sobre um indivíduo singularizado *a priori*, mas pode evidentemente ajudar-nos a conceptualizar o que lhe diz respeito, e tomar decisões com base na avaliação dos riscos.

Suponha-se que se manda fazer a análise a n=1000 indivíduos na situação do Arnesto — isto é, 1000 indivíduos para quem se requer a análise por a sintomatologia, em 30% dos casos, ser causada por porfíria. Como adiante se verá, esta base é totalmente irrelevante, apenas facilita os cálculos — no fim, apenas estamos interessados no *valor preditivo positivo* da análise, que é a fracção $\mathrm{P}(D\,|+) = \dfrac{PV}{P}$.

Uma vez que a prevalência para esta subpopulação de suspeitos é 30%, podemos preencher a linha de totais com os valores $D = 300$, $\overline{D} = 700$.

Se a sensibilidade da análise é 82%, o número de positivos verdadeiros que esperamos encontrar é *PV*=300x0.82=246, e consequentemente o número de negativos falsos é *NF*=300-246=54. E se a especificidade é 96.3%, o número de negativos verdadeiros que esperamos encontrar é *NV*=700x0.963=674.1, donde *PF*=700-674.1=25.9.

Temos assim

	doente	são	total
positivo	246	25.9	271.9
negativo	54	674.1	728.1
total	300	700	1000

Consequentemente, nesta situação, o valor preditivo positivo da análise é

$$\mathrm{P}(D\,|+) = \frac{PV}{P} = \frac{246}{246+25.9} = 90\%.$$

(E o valor preditivo negativo da análise é $\mathrm{P}(\overline{D}\,|-) = \dfrac{NV}{NV+NF} = \dfrac{674.1}{728.1} = 93\%$.)

Nas expressões acima, avaliamos $\mathrm{P}(D)$ pela prevalência da doença *na situação em estudo*, 30% neste caso. Note que esta probabilidade *a priori* se transforma, devido à informação corroborante de que o resultado da análise é positivo, numa probabilidade *a posteriori* de 90%.

Caso II: Suponhamos agora que o médico, confrontado com um resultado positivo da análise à diaminase de porfibilogénio, procede a uma investigação mais cuidada, e confirma-se que o Arnesto tem porfíria. O médico descobre também que o Arnesto tem um irmão, o Bicente, e aconselha-o a fazer a análise. O resultado é positivo. Qual é a probabilidade de o Bicente ter porfíria?

(Admite-se que o Bicente não é gémeo monozigótico do Arnesto; nesse caso seria, evidentemente, portador da doença.)

Como é um carácter dominante, e a prevalência na população é muito baixa (10^{-4}), podemos considerar praticamente nula a probabilidade de ambos os membros de um casal transmitirem em simultâneo esta doença (1 em cada cem milhões de casos), bem como a probabilidade de um dos progenitores transmitir de certeza a doença, pois teria que ser filho de pai e mãe doentes. Assim o progenitor-transmissor pode, de acordo com as considerações atrás feitas, ter transmitido o gene errado com probabilidade 1/2 a cada um dos seus filhos, em particular ao Bicente.

Supondo que se prepara uma tabela como as anteriores, para reflectir no valor preditivo positivo da análise no caso de o indivíduo ter um irmão com porfíria (e partindo mais uma vez da base arbitrária n=1000), isto é, *a priori* $\mathrm{P}(D) = 0.5$, neste caso:

	doente	são	total
positivo	410	18.5	428.5
negativo	90	481.5	571.5
total	500	500	1000

O valor preditivo positivo é

$\mathrm{P}(D\,|+) = \frac{PV}{P} = \frac{410}{428.5} = 96\%$, e o valor preditivo negativo é $\mathrm{P}(\overline{D}\,|-) = \frac{481.5}{571.5} = 84\%$.

Mais uma vez note que não se pode ter o melhor de dois mundos: o valor preditivo positivo aumentou, relativamente ao caso anterior, mas compensatoriamente o valor preditivo negativo baixou!

Caso III: A Tia Elvira[2] ouviu falar da doença, e quis logo ir fazer a análise à diaminase de porfibilogénio.

Hipocondríaca vista por Pedro Veliça
http://anatomias.mediasmile.net[3].

Quando recebeu o resultado foi à pressa para casa pôr a chaleira a ferver para com o vapor de água abrir o envelope, e leu no relatório que o resultado é positivo.

Qual é a probabilidade de a Elvira ter porfíria?

Durante um rastreio da população faz-se a análise a 1000000 de indivíduos (mais uma vez uma base totalmente irrelevante para os resultados, e muito conveniente para os cálculos). Que fazer aos que são considerados positivos?

Como nesta situação estimamos a prevalência da doença na população em 1/10000, podemos preencher a linha de totais com os valores $D = 100$, $\overline{D} = S = 999900$.

Usando como anteriormente os conhecimentos sobre a sensibilidade e a especificidade da análise, obtém-se a tabela

	doente	são	total
positivo	82	36996.3	37078.3
negativo	18	962903.7	962921.7
total	100	999900	1000000

Consequentemente, nesta situação, o valor preditivo positivo da análise é $p = \frac{82}{37078.3} = 0.0022$ (0.22%), ou seja, cerca de 1 em 450 positivos tem de facto a doença!

À primeira vista a análise até parece inútil, tão baixa é a probabilidade de um positivo ter de facto a doença. Mas veja o caso nesta perspectiva: a probabilidade de ter a doença sobe de 1 em 10 000 para 22 em 10 000, no caso de a análise ter resultado positivo!

Por outro lado, o valor preditivo negativo da análise é $q = \frac{962903.7}{962921.7} = 99.998\%$.

Como se vê há algum perigo em tentar um autodiagnóstico a partir dos relatórios de análises. As análises clínicas têm muitos pontos em comum com os testes de hipóteses em Estatística, em particular serem de interpretação delicada, e muitas vezes abusivamente interpretados pelos leigos. Usando a linguagem das análises clínicas como metáfora, quem não perceba bem as distinções pode confundir positivos e doentes, quando são de facto realidades distintas. O leitor pode encontrar informação complementar em Motulsky (2010), ou em Pestana e Velosa (2008, p. 284-289); neste último aborda-se a questão de análises repetidas, confirmatórias ou para contra-prova.

Note que tal como maior sensibilidade arrasta menor especificidade, quando cresce o valor preditivo positivo diminui o valor preditivo negativo, e reciprocamente. Também nos testes de hipóteses quando exigimos menor probabilidade de um "erro de primeira espécie", que consiste em rejeitar uma hipótese nula verdadeira, imediatamente cresce o erro de segunda espécie, que é o de manter uma hipótese nula quando a que consideramos alternativa é verdadeira. Não se pode ter o melhor de dois mundos, há que fazer opções.

2. O teorema de Bayes

O estudo dos três casos foi feito, na Secção 1, usando a linguagem e notações de *tabelas de contingência*, uma das mais antigas ferramentas dos investigadores de todas as áreas (já no século XII Roger Bacon preconizava a análise de "tabelas de presença e de ausência" como excelentes auxiliares de raciocínio). A análise pode ser feita usando um outro instrumento, mais sofisticado, o teorema de Bayes.

Da definição de probabilidade condicional

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

tira-se imediatamente

$$P(A \cap B) = \begin{cases} P(A \mid B) \times P(B) \\ P(B \mid A) \times P(A) \end{cases}$$

e consequentemente

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A) \times \mathrm{P}(A)}{\mathrm{P}(B)}.$$

Temos assim um instrumento que nos permite pesquisar a "probabilidade inversa", isto é, observando as consequências, diagnosticar as suas causas (no sentido de calcular as suas probabilidades, por forma a poder decidir qual a causa mais provável). Evidentemente, como ou ocorre *A* ou o seu complementar Ā, podemos calcular a probabilidade do condicionante *B* como

P (B) = P (B∩A)∪ P (B∩Ā) =
=P (B|A) ×P (A) + P (B|Ā) ×P (Ā) ,

a forma (não trivial) mais simples do teorema da probabilidade total, cuja formulação geral é

$$\mathrm{P}(B) = \sum_{k \in K} \mathrm{P}(B \mid A_k) \times \mathrm{P}(A_k)$$

desde que $\{A_k\}_{k \in K}$ seja uma partição do universo.

Assim, na linguagem da probabilidade condicional e do teorema de Bayes, a análise do caso I da Secção anterior seria

$$\begin{aligned} \text{sensibilidade} &= \mathbb{P}(+ \mid D); \\ \text{especificidade} &= \mathbb{P}(- \mid S); \\ \text{valor preditivo positivo} &= \mathbb{P}(D \mid +); \\ \text{valor preditivo negativo} &= \mathbb{P}(S \mid -). \end{aligned}$$

e portanto

$$\begin{aligned} p &= \mathbb{P}(D \mid +) = \frac{\mathbb{P}(+ \mid D)\ \mathbb{P}(D)}{\mathbb{P}(+ \mid D)\ \mathbb{P}(D) + \mathbb{P}\left(+ \mid \overline{D}\right)\ \mathbb{P}\left(\overline{D}\right)} = \\ &= \frac{0.82 \times 0.3}{0.82 \times 0.3 + (1 - 0.963) \times 0.7} \simeq 0.9, \end{aligned}$$

e

$$\begin{aligned} q &= \mathbb{P}(S \mid -) = \frac{\mathbb{P}(- \mid S)\ \mathbb{P}(S)}{\mathbb{P}(- \mid S)\ \mathbb{P}(S) + \mathbb{P}\left(- \mid \overline{S}\right)\ \mathbb{P}\left(\overline{S}\right)} = \\ &= \frac{0.963 \times 0.7}{0.963 \times 0.7 + (1 - 0.82) \times 0.3} \simeq 0.93. \end{aligned}$$

O Teorema da Probabilidade Total e o seu corolário geralmente designado como Teorema de Bayes são um dos instrumentos mais sofisticados para analisar questões de condicionamento, e é a base de uma das melhores estratégias de amostragem, a amostragem estratificada. São, inclusivamente, a base de uma apologia do raciocínio indutivo na construção do conhecimento, veja-se a discussão das "urnas de Laplace" em Pestana e Velosa (2008, p. 258-265).

Anote-se que o teorema de Bayes é muitas vezes referido como "teorema da probabilidade inversa" (por inverter o sentido do condicionamento), e por "teorema da probabilidade das causas", por ser usado em aplicações médicas para hierarquizar as probabilidades das possíveis causas dos sintomas observados. Nesse sentido, é persistentemente usado quando se procuram relações causais — e por isso mesmo sujeito a controvérsias que parecem intermináveis. A causalidade, longe de ser uma noção ultrapassada como Pearson no seu entusiasmo juvenil advogou (e note-se que *The Grammar of Science* foi aconselhado por Einstein aos seus amigos e discípulos, e mereceu a aprovação de espíritos tão críticos quanto Galton), está na ordem do dia, veja-se o livro notável de Pearl (2009), que estabelece em bases firmes a reflexão sobre causalidade. Veja-se também, por outro lado, na excepcional meditação de Jaynes sobre a lógica da inferência (publicada postumamente, Jaynes and Breethorst, 2003), a discussão dos limites da possibilidade de uma correspondência Física para o que todos os matemáticos fazem quotidianamente — inverter expressões matemáticas — nomeadamente porque o tempo não é reversível.

3. Curvas *ROC* e escolha do ponto de corte

No caso da porfíria, afirmámos que a análise clínica distingue positivos de negativos consoante o nível sérico de diaminase de profibilogénio é inferior ou superior a 99 unidades por mm^3; esta análise tem uma sensibilidade de 82%, e uma especificidade de 93.6%. Como se decide aquele ponto de corte, e se estabelecem as características sensibilidade

$s = \mathrm{P}(+|D)$, a probabilidade de detectar doentes, positivos verdadeiros, e especificidade $e = \mathrm{P}(-|\overline{D})$ a probabilidade de não incomodar os não doentes, negativos verdadeiros?

O desenvolvimento de uma análise clínica é um procedimento demorado, em parte porque tem que haver um tratamento estatístico complexo dos dados disponíveis para estabelecer, exactamente, onde deve ser colocada a fasquia que separa negativos de positivos, e quais as probabilidades $1-s = \mathrm{P}(-|D)$ e $1-e = \mathrm{P}(+|\overline{D})$ de erro, de um e de outro tipo, àquele nível.

Habitualmente há estudos diversos, por diferentes grupos de investigadores, em populações heterogéneas, e usando protocolos desiguais, o que obriga a um trabalho rigoroso de harmonização prévia, ou de síntese (meta-análise) que procura aproveitar esse conhecimento disperso e difuso (Pestana *et al.*, 2006).

Experimentadores diferentes podem pôr mais ou menos ênfase na sensibilidade $s = \mathrm{P}(+|D)$ ou na especificidade $e = \mathrm{P}(-|\overline{D})$ dos testes de diagnóstico, isto é escolher "pontos de corte" diversos. Como construir uma curva *SROC* ("*Summary Receiver Operating Characteristic*")[4] a partir de pontos (1-*e* , *s*) obtidos em diversos estudos?

Comecemos por exemplificar, para clarificar ideias, o caso simples da construção da curva *ROC* no caso da análise ao teor sérico da diaminase de porfibilogénio para diagnóstico da porfíria. Suponha-se que, antes de ter sido acordado que o teste seria feito colocando a fasquia em 99, se estudava o sangue de 53 doentes e de 100 não doentes a diversos níveis, e se registava

• o número *x* de não doentes com resultado negativo, por terem um nível sérico superior ao patamar *T*,

• o número *y* de doentes detectados, isto é com resultado positivo, por terem um nível sérico inferior ao patamar *T*,

por exemplo

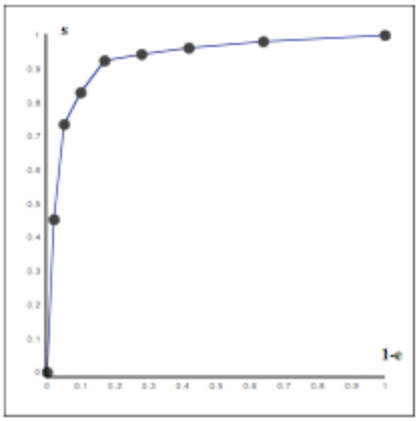
T	80	85	90	95	100	105	110	115	120
x	0	2	5	10	17	28	42	64	100
y	0	24	39	44	49	50	51	52	53

As correspondentes frequências relativas em cada linha dar-nos-iam então as coordenadas (1-*e* , *s*)

T	80	85	90	95	100	105	110	115	120
$1-e$	0.000	0.020	0.050	0.100	0.170	0.280	0.420	0.640	1.000
s	0.000	0.453	0.736	0.830	0.925	0.943	0.962	0.981	1.000

da curva *ROC* empírica (Figura 4).

Figura 4 - Curva *ROC* empírica.



Adiante, num caso concreto, indica-se como se constrói uma curva alisada a partir de algumas coordenadas.

Evidentemente as decisões da Organização Mundial de Saúde, ou de entidades nacionais como o Infarmed, não devem basear-se apenas na evidência de um estudo, feito em geral com recursos limitados. Mas a simples aglomeração de dados de diversos estudos pode dar resultados absurdos, devido a confundimento. Por exemplo, suponha-se que num dos estudos se tem

diagnóstico \estado	D – doente	$\overline{D}$ – não doente
positivo	200	10
negativo	300	60

Tem-se sensibilidade *s* = 200/(200+300) = 0.4, especificidade *e* = 60/(10+60) = 0.86, e o *OR* = (200/10)/(300/60) = 4 indica que esta análise tem um poder discriminativo razoável (*OR* = 1 é equivalente a *s* = 1-*e*, pelo que consideramos *OR* ≈ 1 um indicador de escasso poder discriminativo).

Noutro estudo, os dados são

diagnóstico \estado	D – doente	$\overline{D}$ – não doente
positivo	20	200
negativo	5	200

Tem-se sensibilidade *s* = 20/(20+5) = 0.8, especificidade *e* = 200/(200+200) = 0.5, e o *odds ratio OR* = (20/200)/(5/200) = 4, indicando que também esta análise tem um poder discriminativo razoável.

Porém, a simples aglutinação só dá asneira:

diagnóstico \estado	D – doente	$\overline{D}$ – não doente
positivo	220	210
negativo	305	260

com sensibilidade *s* = 220/(220+305) = 0.42, especificidade *e* = 260/(210+260) = 0.55, e consequentemente *OR* = (220/210)/(305/260) = 0.89, indicando que esta análise tem um poder discriminativo péssimo.

Este efeito perverso da adição dos valores observados em tabelas de contingência é conhecido por paradoxo de Simpson[5]. A forma de meta-analisar tem que ser muito mais sofisticada. Para um bom sumário da sensibilidade/especificidade correspondentes a testes com pontos de corte diversos, faz-se um gráfico de pontos cujas coordenadas são:

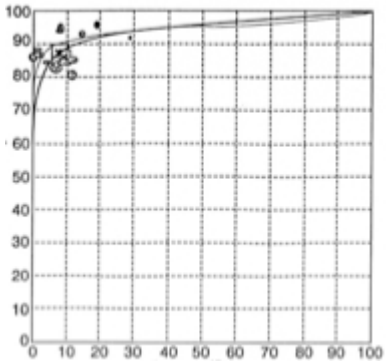
• abcissa — proporção de falsos positivos = $\mathrm{P}(+|\overline{D})$ = 1-*e*;

• ordenada — proporção de positivos verdadeiros = $\mathrm{P}(+|D)$ = s.

É de esperar um crescimento rápido da sensibilidade de 0 a valores próximos de 0.90 quando a abcissa *x* = 1–*e* varia de 0 a 0.10, seguindo-se um crescimento muito lento até 1, como se observa na Figura 5.

Por outras palavras, esperamos que a curva *ROC* esteja consideravelmente acima da bissectriz *s* = 1-*e*.

Figura 5 - Curva ROC em estudos de ultrassons para estenose da artéria carótida.



Se a curva *ROC* passasse pelo ponto (0,1), isso significaria que era possível ter a situação ideal de 100% de sensibilidade e de especificidade. Mas não é possível ter o melhor de dois mundos, é evidente que para aumentar a sensibilidade alguma coisa há que sacrificar na especificidade, e vice-versa. Mas por outro lado também esperamos que a curva *ROC* se afaste visivelmente da bissectriz *s* = 1-*e* ⇔ *OR* = 1, que corresponderia a um teste de diagnóstico muito pouco recomendável, no sentido em que o ganho em sensibilidade de ponto de corte para ponto de corte corresponde exactamente à perda de especificidade.

O *odds ratio OR* pode ser escrito $OR = \frac{LR^+}{LR^-}$, onde LR^+ = *s*/(1-*e*) e LR^- = (1-*s*)/*e* são as "razões de verosimilhanças" que nos permitem, usando a fórmula de Bayes, transformar as vantagens pré-teste em vantagens pós-teste.

Se por outro lado considerarmos como avaliação da assimetria nos ganhos de sensibilidade *versus* perdas de especificidade $LR^+ vs. LR^-$, admitindo que há

assimetria entre o que se ganha numa das características e se perde na outra, ao alterar o ponto de corte (isto é, o teste favorece, para cada ponto de corte, o diagnóstico correcto ou de doentes ou de não doentes), parece razoável procurar uma curva exprimindo *OR* como função de *S*. Espera-se então que uma transformação logarítmica linearize essa relação:

$$\ln OR = \ln LR^{+} - \ln LR^{-} \approx \alpha \ln S + \beta = \\ = \alpha \left(\ln LR^{+} + \ln LR^{-} \right) + \beta$$

equivalente a

$$\ln \frac{a}{b} + \ln \frac{b+d}{a+c} - \ln \frac{c}{d} - \ln \frac{b+d}{a+c} \approx$$

$$\approx \alpha \left(\ln \frac{a}{b} + \ln \frac{b+d}{a+c} + \ln \frac{c}{d} + \ln \frac{b+d}{a+c} \right) + \beta$$

ou ainda

$$\ln \frac{a}{b} - \ln \frac{c}{d} \approx \alpha \left(\ln \frac{a}{b} + \ln \frac{c}{d} \right) + \beta^{*},$$

Onde

$$\beta^{*} = \beta + \ln \left(\frac{b+d}{a+c} \right)^{2\alpha}.$$

Uma forma simples de estimar essa curva *ROC* (Littenberg *et al*., 1990) é então, para cada uma das tabelas 2x2

	D	$\overline{D}$
+	a_k	b_k
−	c_k	d_k

• Calcular as expressões $\hat{U}_k$ e $\hat{V}_k$, ajustados por causa de eventuais zeros,

$$\hat{U}_k = \ln \frac{c_k + 0.5}{d_k + 0.5}$$

e

$$\hat{V}_k = \ln \frac{a_k + 0.5}{b_k + 0.5}.$$

• Ajustar uma recta (ajustamento resistente, cf. Pestana e Velosa, 2008, p. 178-185, ou usando mínimos quadrados ponderados) aos pontos

$$\left(\hat{V}_k + \hat{U}_k, \hat{V}_k - \hat{U}_k \right).$$

• Deduzir daquela recta de regressão a expressão analítica da curva *ROC*, como exemplificado na Figura 5.

Note-se que isto corresponde a admitir uma forma muito genérica para as curvas *ROC*,

$$\left(\frac{s}{1-e} \right)^{1-\alpha} \left(\frac{e}{1-s} \right)^{1+\alpha} =$$

$$= \frac{e^{\alpha+1} (1-e)^{\alpha-1}}{s^{\alpha-1} (1-s)^{\alpha+1}} = C = \exp(\beta).$$

Se não houver condicionantes extra, parece natural definir o ponto de corte óptimo, no sentido em que dá um equilíbrio apelativo no que refere sensibilidade e especificidade, com base no ponto mais próximo do óptimo ideal — inatingível — (0,1). Por outras palavras, trata-se de minimizar o quadrado da distância $d^2=(1-s)^2+(1-e)^2$ de um ponto genérico $(1-e,s)$ ao canto superior esquerdo do quadrado $\{(x,y): 0 \leq x \leq 1, 0 \leq y \leq 1\}$.

Mas, evidentemente, é em geral mais sensato considerar penalizações definindo funções de perda L_1 e L_2 associadas aos diagnósticos errados, diagnosticar negativo um doente ou diagnosticar positivo um não doente, respectivamente, e o problema efectivo é procurar o ponto da curva *ROC* ajustada que minimiza a soma L_1+L_2 das perdas. Veja-se também van Belle (2002, secção 4.10), que aborda com detalhe e pragmatismo a construção de uma curva *ROC*.

Um foco de preocupação é se a curva *ROC* assim estimada denuncia heterogeneidade no que refere patamares de decisão (*thresholds*), ou na capacidade discriminativa, ou em ambos. Deve procurar explicar-se essa heterogeneidade, indicar ao menos se deriva de características específicas diversas dos testes de diagnóstico que estamos a tentar harmonizar, de diferenças populacionais, ou de planeamentos experimentais diversos.

Notas

1. Investigação parcialmente financiada por FCT/OE.
2. Este enunciado é uma homenagem à tia Elvira, hipocondríaca extreme que viveu cheia de saúde até aos 92 anos, apesar de ter todas as doenças de que ouvia falar — quando alguém tinha uma doença perguntava logo os sintomas, para conferir, e não é que conferia sempre! Só num almoço de Natal é que um dos netos a conseguiu entupir: estava ela num despique de "eu sou mais doente do que tu" com uma outra hipocondríaca, cada qual com as suas armas: "Veja estas minhas análises, D. Elvira, como está tudo descontrolado, eu estou muito doente", ao que a tia Elvira ripostava indo buscar radiografias e exclamando "E esta minha vesícula, veja aqui, Paula, nesta radiografia quase não aparece, e naquela tão volumosa! Não admira que me sinta tão mal". Esta foi a deixa aproveitada pelo Carlos: Foi deixando cair que tinha que fazer uma ecografia porque tinha sintomas preocupantes, depois de muito instado lá disse que era ao pénis, e quando a avó disse muito irritada que ele estava sempre a brincar com coisas sérias (a saúde dela, com certeza), ele explicou que tinha nesse órgão sensível os mesmos sintomas preocupantes que a avó tinha na vesícula: umas vezes estava volumoso, outras vezes não …
3. Agradeço ao Pedro Veliça a autorização para a usar a *Anatomia da Hipocondríaca*, retirada de: http://anatomias.mediasmile.net/
4. Por simplicidade, vamos a partir de agora falar de curvas *ROC*, sendo claro do contexto que quando usamos pontos obtidos em diversos estudos se trata de uma síntese, e portanto de uma curva *SROC*.
5. Deveria de facto ser denominado paradoxo de Yule (1903), e não de Simpson (1951), cf. David and Edwards, 2001, p. 137-140.

REFERÊNCIAS

David, H. A., and Edwards, A. W. F. (2001). *Annotated Readings in the History of Statistics*, Springer, New York.

Fisher, R. A. (1995). *Statistical Methods, Experimental Design and Scientific Inference* (reedição conjunta de *Statistical Methods for Research Workers*, *The Design of Experiments*, e *Statistical Methods and Scientific Inference*), Oxford Univ. Press, Oxford.

Gigerenzer, G. (2002). *Calculated Risks: How to Know When Numbers Deceive You*, Simon and Schuster, New York.

Goldacre, B. (2008). *Bad Science*, Fourth Estate (trad.: *Ciência da Treta*, Bizâncio, Lisboa, 2010).

Jaynes, E. T., and Bretthorst, G. L. (2003). Probability Theory: The Logic of Science, Cambridge University Press, Cambridge.

Littenberg, B., Moses, L., and Rabinowitz, D. (1990). Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method, *Clinical Research* 38, 415A.

Motulsky, H. (2010). *Intuitive Biostatistics: A Mathematical Guide to Statistical Thinking*, Oxford University Press, Oxford.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, 2nd ed., Cambridge University Press, Cambridge.

Pearson, K. (2009). The Grammar of Science, Bibliobazar.

Pestana, D., Velosa, S., Sequeira, F., e Vasconcelos, R. (2006). Evidência Estatística e Meta-Análise, in G. Cunha e J. Varanda (eds.), *Estatística e Qualidade na Saúde*, Lisboa, 48-74.

Pestana, D., e Velosa, S. (2008). *Introdução à Probabilidade e à Estatística*, vol. I, 3ª ed., Fundação Gulbenkian, Lisboa.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables, *J. Roy. Statist. Soc.* B13, 238-241.

van Belle, G. (2002). *Statistical Rules of Thumb*, Wiley, New York.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics, *Biometrika* 2, 121-134.