ANÁLISE BAYESIANA DO DESEMPENHO DE DOIS TESTES DIAGNÓSTICOS QUANDO INDIVÍDUOS COM RESULTADOS NEGATIVOS EM AMBOS OS TESTES NÃO SÃO VERIFICADOS POR UM PADRÃO-OURO

Edson Zangiacomi MARTINEZ[1]
Jorge Alberto ACHCAR[2]
Francisco LOUZADA-NETO[2]

- RESUMO: Na literatura médica, os estudos sobre o desempenho de testes diagnósticos utilizam freqüentemente as medidas de sensibilidade e especificidade. Essas medidas estimam, respectivamente, a probabilidade de um teste fornecer um resultado positivo, dado que o indivíduo realmente é portador da doença, e a probabilidade do teste fornecer um resultado negativo, dado que o indivíduo não é portador da doença. Essas medidas são obtidas da comparação direta dos resultados do teste e de um procedimento denominado padrão-ouro, que classifica os indivíduos corretamente como doentes e não doentes. No entanto, em alguns estudos, são submetidos ao padrão-ouro somente os indivíduos com resultados positivos em pelo menos um dos testes sob investigação. No presente estudo, propomos o uso de um modelo Bayesiano para a estimação da sensibilidade e da especificidade nesta situação. Uma extensão do modelo é apresentada, de modo que o efeito de co-variáveis sobre as medidas de desempenho possa ser estudado. Uma aplicação a dados reais, obtidos da literatura médica, ilustra a metodologia.

- PALAVRAS CHAVE: Sensibilidade; especificidade; testes diagnósticos; métodos Bayesianos.

1 Introdução

Os métodos estatísticos aplicados à medicina diagnóstica apresentaram enormes avanços nas últimas décadas. Grande parte desses métodos está voltada ao problema de classificar indivíduos em grupos, os testes diagnósticos compõem o principal exemplo. Esses testes são descritos como métodos capazes de indicar a presença ou a ausência de uma determinada doença, com uma certa chance de erro, e a quantificação dessas chances de erro é basicamente o objetivo dos métodos estatísticos. Na terminologia médica, a probabilidade de um teste diagnóstico produzir um resultado positivo, dado que o indivíduo é realmente portador da doença, é chamada sensibilidade (*S*) do teste, ou taxa de verdadeiros positivos (*TVP*). A probabilidade do teste produzir um resultado negativo,

[1]Departamento de Medicina Social, Faculdade de Medicina de Ribeirão Preto - FMRP, Universidade de São Paulo – USP, CEP 14049-900, Ribeirão Preto, SP, Brasil. E-mail: *edson@fmrp.usp.br.*

[2]Departamento de Estatística, Universidade Federal de São Carlos – UFSCar, Caixa Postal 676, CEP 13565-905, São Carlos, SP, Brasil. E-mail: jachcar@power.ufscar.br / dfln@power.ufscar.br.

dado que o indivíduo não porta a doença, é chamada especificidade (E). O complemento de E, ou seja, $1 - E$, é freqüentemente chamado de taxa de falsos positivos (TFP). Essas medidas são tradicionalmente obtidas da comparação dos resultados dos testes com os resultados de um teste de referência, denominado genericamente padrão-ouro.

Em linguagem matemática, a sensibilidade de um teste diagnóstico é dada por $S = P(T=1|D=1)$ e a especificidade é dada por $E = P(T=0|D=0)$, onde $T = 1$ representa um resultado positivo para o teste, $T=0$ representa um resultado negativo, e $D = 1$ e $D = 0$ representam o verdadeiro estado do indivíduo segundo um padrão-ouro, respectivamente, doente ou não doente. Um padrão ouro ideal classifica corretamente os indivíduos como doentes ou não doentes, mas, na prática, são geralmente procedimentos invasivos, caros, muitas vezes dolorosos, que podem oferecer riscos ao paciente, e usualmente não são aplicados a indivíduos com pouca evidência de serem portadores da doença (Sox, 1986). Ilustrando esta situação, Claeys et al. (2003) conduziram um estudo objetivando estimar o desempenho da inspeção visual com ácido acético e da citologia cervical (teste de Papanicolaou) em detectar lesões precursoras do câncer do colo do útero, sendo que todas as mulheres que compuseram a amostra foram submetidas a ambos os testes, mas apenas as que receberam resultado positivo em pelo menos um dos testes foram submetidas ao padrão-ouro. O teste de referência escolhido foi a colposcopia seguida de biópsia, um procedimento não indicado a mulheres sem evidências da doença.

Um método estatístico dirigido à análise de dados de medicina diagnóstica em que os indivíduos com resultados negativos em todos os testes sob investigação não são verificados pelo padrão ouro foi inicialmente proposto por Schatzkin et al. (1987). Este método trata especificamente da investigação do desempenho de dois testes diagnósticos, e tem por limitação não estimar diretamente sensibilidades e especificidades, mas sim o quanto um teste tem melhor desempenho que outro. Sejam dois testes diagnósticos denotados respectivamente por T_1 e T_2, onde $T_v = 1$ representa um resultado positivo para o teste v, $v = 1,2$, e $T_v = 0$ representa um resultado negativo. A Tabela 1 representa, genericamente, os resultados de ambos os testes perante um padrão-ouro ideal. Os valores d e h que aparecem entre colchetes são desconhecidos, embora a soma $d + h$, ou seja, o total de indivíduos não verificados pelo padrão-ouro, seja conhecida. Os valores n_+ e n_- também são desconhecidos, e referem-se aos totais de indivíduos classificados respectivamente como doentes e não doentes pelo padrão-ouro. Por notação, S_v e E_v são respectivamente as sensibilidades e especificidades do teste v, v=1,2.

Tabela 1 - Representação genérica dos resultados de dois testes diagnósticos, em que os indivíduos negativos em ambos testes não são verificados pelo padrão-ouro. Os valores entre colchetes são desconhecidos

	Doentes			Não doentes		
	$T_2 = 1$	$T_2 = 0$	Total	$T_2 = 1$	$T_2 = 0$	Total
$T_1 = 1$	a	b	$a + b$	e	f	$e + f$
$T_1 = 0$	c	$[d]$	$c + [d]$	g	$[h]$	$g + [h]$
Total	$a + c$	$b + [d]$	$[n_+]$	$e + g$	$f + [h]$	$[n_-]$

Schatzkin et al. (1987) mostraram que estimativas das sensibilidades dos testes 1 e 2, dadas respectivamente por $(a+b)/n_+$ e por $(a+c)/n_+$, não podem ser diretamente obtidas,

pois n_+ é desconhecido. Por outro lado, a razão S_1/S_2 é estimada diretamente por $(a+b)/(a+c)$. Essa razão é denominada por Schatzkin et al. (1987) de taxa relativa de verdadeiros positivos (*rTVP*), e é interpretável como o número de vezes em que a sensibilidade do teste 1 é maior que a sensibilidade do teste 2. Por sua vez, a razão de especificidades E_1/E_2 não pode ser analogamente estimada por $(g+h)/(f+h)$, pois h é um valor desconhecido. Como alternativa, Schatzkin et al. (1987) propôs a estimação da razão $(1 - E_1)/ (1 - E_2)$, denominada taxa relativa de falsos positivos (*rTFP*). Um estimador de *rTFP* é dado simplesmente por $(e+f)/(e+g)$.

A construção de intervalos de confiança para *rTVP* e para *rTFP* é discutida por Cheng e Macaluso (1997), Chock et al. (1997), Cheng et al. (2000) e, mais recentemente, por Alonzo et al. (2004).

Walter (1999) propôs um modelo capaz de estimar as sensibilidades e as especificidades de dois testes diagnósticos quando os indivíduos com resultados negativos em ambos testes não são submetidos ao padrão-ouro. Uma extensão Bayesiana desse modelo foi posteriormente proposta por Van der Merwe e Maritz (2002), utilizando amostradores de Gibbs. No presente artigo, apresentamos um modelo Bayesiano bastante similar ao proposto por Van der Merwe e Maritz (2002) e por Joseph et al. (1995) e estendemos a proposta à situação em que um vetor de co-variáveis está presente. Nesta modelagem, utilizamos o algoritmo de Metropolis-Hastings. Para ilustrar uma aplicação do modelo proposto a dados reais, utilizamos dados retirados da literatura médica, na qual obtivemos estimativas das medidas de sensibilidade, especificidade e prevalência.

2 Formulação do modelo

Joseph et al. (1995) introduziram um modelo Bayesiano baseado em amostradores de Gibbs para estimar as medidas de desempenho de um ou mais testes diagnósticos em situações em que um padrão-ouro não é disponível. Esse modelo pode ser facilmente estendido às situações em que os indivíduos amostrados com resultados negativos aos testes diagnósticos sob investigação não são verificados pelo padrão-ouro. Seja p a prevalência da doença em questão na população e D a verdadeira condição do indivíduo, onde D=1 denota um indivíduo com a doença e D=0 denota um indivíduo livre da doença. Assim, $p = P(D=1)$. Sejam T_1 e T_2 duas variáveis aleatórias relativas aos resultados dos testes diagnósticos, onde $T_v = 1$ denota um teste com resultado positivo e $T_v = 0$ denota um resultado negativo, para v=1,2. As sensibilidades dos testes diagnósticos são dadas por $S_v = P(T_v=1|D=1)$ e as especificidades dadas por $E_v=P(T_v=0|D=0)$, onde v=1,2. Da aplicação da teoria da probabilidade, podemos escrever a relação:

$$P(T_1=1,\ T_2=1,\ D=1) = P(D=1)\ P(T_1=1|\ D=1)\ P(T_2=1|\ T_1=1,\ D=1). \qquad (1)$$

Sob o pressuposto de que T_1 e T_2 são variáveis aleatórias independentes, a equação (1) pode ser reescrita da forma:

$$P(T_1=1,\ T_2=1,\ D=1) = P(D=1)\ P(T_1=1|\ D=1)\ P(T_2=1|\ D=1) = pS_1S_2. \qquad (2)$$

Analogamente, podemos obter as probabilidades associadas às outras sete possíveis combinações de D e dos resultados dos testes T_1 e T_2, como exibido na Tabela 2. O produto dessas probabilidades, elevadas às respectivas quantidades mostradas na quarta

coluna da Tabela 2, define a função de verossimilhança $L(\theta_1)$, onde θ_1 é o vetor de parâmetros $\theta_1^T = (S_1, S_2, E_1, E_2, p)$.

Tabela 2 - Contribuições de todas as possíveis combinações entre as variáveis D, T_1 e T_2 à verossimilhança. Os valores entre colchetes são desconhecidos

D	T_1	T_2	Número de indivíduos	Contribuição à verossimilhança
1	1	1	a	pS_1S_2
1	1	0	b	$pS_1(1-S_2)$
1	0	1	c	$p(1-S_1)S_2$
1	0	0	[d]	$p(1-S_1)(1-S_2)$
0	1	1	e	$(1-p)(1-E_1)(1-E_2)$
0	1	0	f	$(1-p)(1-E_1)E_2$
0	0	1	g	$(1-p)E_1(1-E_2)$
0	0	0	[h]	$(1-p)E_1E_2$

Na Tabela 2, d e h são valores desconhecidos, mas $u = d + h$ é uma quantidade conhecida, dada pelo número total de indivíduos não verificados. Para não haver confusão na notação entre quantidades conhecidas e desconhecidas, sejam $Y_1 = d$ e $Y_2 = h$.

As variáveis Y_1 e Y_2 referem-se respectivamente ao número de indivíduos doentes não verificados e ao número de indivíduos não doentes e não verificados. Estas variáveis, na terminologia introduzida por Tanner e Wong (1987), são definidas como variáveis latentes, desde que existe a possibilidade de reproduzi-las por meio de suas distribuições de probabilidade, apesar de não serem observáveis. A partir da fórmula de Bayes, verifica-se que a distribuição condicional de Y_1 é dada por:

$$Y_1 \mid u, \theta_1 \sim \text{binomial}\left(u; \frac{p(1-S_1)(1-S_2)}{p(1-S_1)(1-S_2)+(1-p)E_1E_2}\right), \tag{3}$$

e $Y_2 = u - Y_1$ (lembrando que Y_1 e Y_2 são desconhecidos, mas u é conhecido). Na análise Bayesiana do modelo é assumido que as densidades *a priori* para todos os parâmetros do vetor θ_1 possuem distribuição beta, onde α_θ e β_θ genericamente denotam os hiperparâmetros (conhecidos) da distribuição de θ. Sendo $Z^T=(a,b,c,e,f,g)$ um vetor de quantidades observáveis, e combinando a função de verossimilhança com as densidades *a priori*, as demais distribuições condicionais para o algoritmo de amostradores de Gibbs são dadas por:

$$\begin{aligned} p \mid Z, Y_1, Y_2, \alpha_p, \beta_p &\sim \text{Beta}\,(a+b+c+Y_1+\alpha_p\,;\, e+f+g+Y_2+\beta_p), \\ S_1 \mid Z, Y_1, Y_2, \alpha_{S_1}, \beta_{S_1} &\sim \text{Beta}\,(a+b+\alpha_{S_1}\,;\, c+Y_1+\beta_{S_1}), \end{aligned} \tag{4}$$

$$S_2 \mid \mathbf{Z},\ Y_1,\ Y_2,\ \alpha_{S_2},\ \beta_{S_2} \sim \text{Beta}\ (a+c+\alpha_{S_2}\ ;\ b+Y_1+\beta_{S_2}\ ),$$

$$E_1 \mid \mathbf{Z},\ Y_1,\ Y_2,\ \alpha_{E_1},\ \beta_{E_1} \sim \text{Beta}\ (g+\ Y_2+\alpha_{E_1}\ ;\ e+f+\beta_{E_1}\ ),\quad \text{e}$$

$$E_2 \mid \mathbf{Z},\ Y_1,\ Y_2,\ \alpha_{E_2},\ \beta_{E_2} \sim \text{Beta}\ (f+\ Y_2+\alpha_{E_2}\ ;\ e+g+\beta_{E_2}\ ).$$

O primeiro passo do algoritmo consiste em gerar aleatoriamente Y_1 de (3), dados valores arbitrários para os parâmetros do vetor θ_1. A seguir, utilizando o valor Y_1 obtido, observações de p, S_1, S_2, E_1 e E_2 são geradas aleatoriamente conforme suas respectivas distribuições condicionais (4). Os valores de p, S_1, S_2, E_1 e E_2 assim obtidos são usados para gerar aleatoriamente um novo valor de Y_1, que, por sua vez, é utilizado para gerar novos valores de p, S_1, S_2, E_1 e E_2, e este ciclo é assim repetido um número grande de vezes. As amostras aleatórias geradas são usadas para estimar as densidades marginais *a posteriori* de cada parâmetro em θ_1 e das variáveis Y_1 e Y_2.

3 Presença de co-variáveis

Seja um vetor de co-variáveis X_i, i=1,...,n, e sejam p, S_1, S_2, E_1 e E_2 relacionados a X_i da forma:

$$\theta_{ki} = \frac{\exp(\sum_{j=0}^{L} \gamma_{kj} x_{ji})}{1+\exp(\sum_{j=0}^{L} \gamma_{kj} x_{ji})}, \tag{5}$$

onde $X_{0i} = 1$, i = 1,...,n, k=1,...,5, L é o número de co-variáveis e θ_{ki} genericamente denota um parâmetro de interesse, ou seja, $\theta_{1i} = S_{1_i}$, $\theta_{2i} = S_{2_i}$,$\theta_{3i} = E_{1_i}$,$\theta_{4i} = E_{2_i}$ e $\theta_{5i} = p_i$. Definimos assim o vetor de parâmetros $\theta_2^{\mathrm{T}}=(\gamma_1,\gamma_2,\gamma_3,\gamma_4,\gamma_5)$, onde $\gamma_1^{\mathrm{T}}=(\gamma_{10},\gamma_{11},...,\gamma_{1L})$, $\gamma_2^{\mathrm{T}}=(\gamma_{20},\gamma_{21},...,\gamma_{2L})$,..., $\gamma_5^{\mathrm{T}}=(\gamma_{50},\gamma_{51},...,\gamma_{5L})$. Dado T_{1_i} =0 e T_{2_i} =0, i = 1,...,n, seja G_i uma variável aleatória com distribuição de Bernoulli, com probabilidade de sucesso dada por:

$$h_i = \frac{p_i\left(1-S_{1_i}\right)\left(1-S_{2_i}\right)}{p_i\left(1-S_{1_i}\right)\left(1-S_{2_i}\right)\ +\left(1-p_i\right)E_{1_i}E_{2_i}}. \tag{6}$$

Como G_i é uma variável observável quando T_{1_i} =1 ou T_{2_i} =1, ou seja, G_i equivale nesta situação ao verdadeiro e conhecido estado do indivíduo (1 se doente ou 0 se não doente), e não observável quando T_{1_i} =0 e T_{2_i} =0, G_i é uma variável denominada semi-latente, segundo a terminologia de Van der Merwe e Maritz (2002). A função de verossimilhança para θ_2 é dada por:

$$L(\theta_2)=\frac{\exp\left(\sum_{m=1}^{2}\sum_{l=0}^{k}\gamma_{ml}\sum_{i=1}^{n}x_{il}t_{m_i}g_i+\sum_{m=3}^{4}\sum_{l=0}^{k}\gamma_{ml}\sum_{i=1}^{n}x_i\left(1-t_{(m-2)_i}\right)(1-g_i)+\sum_{l=0}^{k}\gamma_{5l}\sum_{i=1}^{n}x_{il}g_i\right)}{\prod_{i=1}^{n}\left\{\prod_{m=1}^{2}\left[1+\exp\left(\sum_{l=0}^{k}\gamma_{ml}x_{il}\right)\right]^{g_i}\prod_{m=3}^{4}\left[1+\exp\left(\sum_{l=0}^{k}\gamma_{m_i}x_{il}\right)\right]^{1-g_i}\left[1+\exp\left(\sum_{l=0}^{k}\gamma_{5l}x_{il}\right)\right]\right\}}, \quad (7)$$

onde g_i é uma observação de G_i. Considerando as densidades *a priori* (independentes) γ_{kj} ~$N(a_{kj};b^2_{kj})$, onde k=1,...,5, j=0,1,...,L, e a_{kj} e b_{kj} são hiperparâmetros conhecidos, e sendo D o conjunto das observações amostrais e dos valores de G_i, as densidades condicionais *a posteriori* são dadas por

$$\pi\left(\gamma_{mj}\mid\boldsymbol{\theta}_{2(\gamma_{mj})},\boldsymbol{D}\right)\propto N(a_{mj};b^2_{mj})\times\exp\left\{\gamma_{mj}\sum_{i=1}^{n}x_{ij}t_{m_i}g_i-\sum_{i=1}^{n}g_i\ln\left[1+\exp\left(\sum_{l=0}^{k}\gamma_{ml}x_{il}\right)\right]\right\} \quad (8)$$

onde m = 1,2 e j = 0,...,L,

$$\pi\left(\gamma_{mj}\mid\boldsymbol{\theta}_{2(\gamma_{mj})},\boldsymbol{D}\right)\propto N(a_{mj};b^2_{mj})\times$$
$$\times\exp\left\{\gamma_{mj}\sum_{i=1}^{n}x_{ij}\left(1-t_{m_i}\right)(1-g_i)-\sum_{i=1}^{n}(1-g_i)\ln\left[1+\exp\left(\sum_{l=0}^{k}\gamma_{ml}x_{il}\right)\right]\right\}, \quad (9)$$

onde m = 3,4 e j = 0,...,L, e:

$$\pi\left(\gamma_{5j}\mid\boldsymbol{\theta}_{2(\gamma_{5j})},\boldsymbol{D}\right)\propto N(a_{5j};b^2_{5j})\times\exp\left\{\gamma_{5j}\sum_{i=1}^{n}x_{ij}g_i-\sum_{i=1}^{n}\ln\left[1+\exp\left(\sum_{l=0}^{k}\gamma_{5l}x_{il}\right)\right]\right\} \quad (10)$$

onde, por exemplo, $\boldsymbol{\theta}_{2(\gamma_{10})}$ é o vetor θ_2 sem o parâmetro γ_{10}. Como as distribuições condicionais (8) a (10) não possuem uma forma conhecida, utilizamos o algoritmo de Metropolis-Hastings (Smith e Roberts, 1993) para simular amostras de γ_{kj}, k=1,...,5, j=0,1,...,L.

4 Um exemplo numérico: resultados de um rastreamento do câncer de próstata

Smith et al. (1997) submeteram 18.527 homens brancos e 949 negros, com 50 anos de idade ou mais, a dois testes utilizados no rastreamento do câncer de próstata: o antígeno prostático específico (PSA, sigla em inglês para *prostate-specific antigen*) e o toque retal (DRE, sigla em inglês para *digital rectal examination*). Concentrações séricas do PSA maiores que 4 ng/ml foram classificadas como positivas. Os indivíduos foram submetidos à biópsia somente se pelo menos um dos testes mostrou resultado sugestivo de câncer de próstata. Os resultados desse estudo são mostrados na Tabela 3.

Tabela 3 - Resultados dos testes PSA e DRE (Smith et al., 1997)

	Com câncer de próstata		Sem câncer de próstata		Todos os indivíduos	
	DRE +	DRE –	DRE +	DRE –	DRE +	DRE –
PSA +	189	292	141	755	330	1.047
PSA –	145	?	1.002	?	1.147	16.952

Das densidades condicionais dadas em (3) e (4), geramos 100.000 amostras por meio do algoritmo de amostradores de Gibbs. Para a geração dessas cadeias desenvolvemos uma rotina computacional na linguagem IML (*Interactive Matrix Language*) do programa SAS versão 8 (SAS Institute Inc., 1999). Das cadeias geradas, descartamos as primeiras 20.000 amostras e tomamos cada qüinquagésima amostra, resultando em um tamanho amostral de 1.600 para as inferências. Os resultados obtidos são mostrados na Tabela 4, em que aparecem as médias *a posteriori* para os parâmetros, com seus respectivos intervalos de credibilidade 95%. Os limites inferior e superior desses intervalos de credibilidade são dados, respectivamente, pelos percentis 2,5% e 97,5% das distribuições *a posteriori* geradas pelo algoritmo. Todos os hiperparâmetros das distribuições *a priori* foram fixados em 0,5, com o objetivo de fixarmos distribuições não informativas (Box e Tiao, 1973). Em uma análise de sensibilidade, chegamos a resultados bastante próximos associando a todos os parâmetros distribuições *a priori* betas com todos os respectivos hiperparâmetros iguais a 1. A convergência foi monitorada graficamente e, por meio do método de Geweke (1992), implementado no programa CODA (Best et al., 1995). Observou-se que as cadeias correspondentes a todos os parâmetros convergiram.

Tabela 4 - Médias *a posteriori* para os parâmetros, com seus respectivos intervalos de credibilidade 95%

			Informação *a posteriori*		
Teste	Parâmetro	Distribuição *a priori*	Média	Intervalo de credibilidade 95%	
PSA	S_1	Beta(0,5;0,5)	0,563	0,510	0,618
	E_1	Beta(0,5;0,5)	0,957	0,954	0,960
DRE	S_2	Beta(0,5;0,5)	0,391	0,347	0,434
	E_2	Beta(0,5;0,5)	0,946	0,942	0,949
	p	Beta(0,5;0,5)	0,039	0,035	0,043

A Tabela 5 compara estimativas clássicas e Bayesianas dos parâmetros *rTVP* e *rTFP*. As estimativas clássicas foram obtidas como explicitado na introdução deste artigo, e seus respectivos intervalos de confiança 95% foram calculados pelo método assintótico delta, como proposto por Cheng e Macaluso (1997). As estimativas Bayesianas são derivadas das amostras de Gibbs geradas para a obtenção das medidas sumarizadas na Tabela 4. Observamos que as estimativas clássicas e Bayesianas são muito próximas, e que os intervalos de confiança e de credibilidade são bem semelhantes.

Tabela 5 - Estimativas clássicas e Bayesianas dos parâmetros *rTVP* e *rTFP*

Parâmetro	Método clássico			Método Bayesiano (média *a posteriori*)		
	Estimativa	Intervalo de confiança 95%		Média	Intervalo de credibilidade 95%	
rTVP	1,440	1,300	1,595	1,443	1,303	1,594
rTFP	0,784	0,722	0,850	0,785	0,718	0,854

Com o objetivo de demonstrar o modelo Bayesiano introduzido na seção 3, que possibilita o estudo do desempenho de testes diagnósticos de acordo com a observação de uma co-variável, usamos ainda os dados de Smith et al. (1997), conforme apresentados na Tabela 6. Observamos nesta Tabela os mesmos dados apresentados na Tabela 3, mas agora divididos em dois grupos distintos, conforme a cor da pele dos indivíduos submetidos aos testes PSA e DRE. A covariável X em questão é, portanto, a cor da pele, onde uma observação x_i de X, i=1,...,n, é rotulada como 0 se o i-ésimo indivíduo é branco, e como 1 se o i-ésimo indivíduo é negro. Temos portanto um modelo com 10 parâmetros, γ_{kj}, onde k=1,...,5 e j=0,1. Os hiperparâmetros a_{kj}, k=1,...,5, j=0,1, foram escolhidos de acordo com os resultados do ajuste de dois modelos preliminares sem co-variáveis (como descrito na seção 2), sendo o primeiro baseado somente na amostra de homens negros e o segundo baseado somente na amostra de homens brancos. Estamos dessa maneira adotando um método Bayesiano empírico (Carlin e Louis, 2000), em que as distribuições *a priori* são obtidas de uma análise preliminar dos dados. Por sua vez, os hiperparâmetros b_{kj}, k=1,...,5, j=0,1, associados aos desvios-padrão das densidades (independentes) *a priori*, foram fixados em 0,1. Tal escolha também se baseou nos resultados dos modelos preliminares. Das densidades condicionais dadas em (8) a (10), usamos uma outra rotina escrita na linguagem IML do programa SAS para gerar 50.000 amostras por meio do algoritmo de Metropolis-Hastings. Tomamos cada décima amostra, após descartarmos as primeiras 10.000.

Tabela 6 - Resultados dos testes PSA e DRE (Smith et al., 1997), para 949 homens negros e 18.527 homens brancos

		Com câncer de próstata		Sem câncer de próstata		Todos os indivíduos	
		DRE +	DRE –	DRE +	DRE –	DRE +	DRE –
Negros	PSA +	10	28	3	38	13	66
	PSA –	8	?	26	?	34	836
Brancos	PSA +	179	264	138	717	317	981
	PSA –	137	?	976	?	1.113	16.116

A Tabela 7 mostra estimativas Bayesianas dos parâmetros baseadas nas amostras geradas pelo algoritmo. A estimativa do parâmetro γ_{21} é negativa (-0,6477), com um intervalo de credibilidade 95% que não inclui o valor zero, o que sugere que o DRE é

menos sensível para os homens negros do que para os brancos. O parâmetro γ_{41} é estimado em 0,7115, também com um intervalo de credibilidade que não inclui o valor zero, indicando maior especificidade do DRE para os indivíduos negros. O modelo Bayesiano sugere também que a prevalência de câncer prostático é maior para os negros, visto que o parâmetro γ_{51} é estimado em 0,5902, com um intervalo de credibilidade 95% que não inclui o valor zero. De fato, a literatura médica indica que a prevalência de câncer de próstata é geralmente maior em negros do que em brancos (Crawford, 2003).

Tabela 7 - Estimativas Bayesianas para os parâmetros, com seus respectivos intervalos de credibilidade 95%

	Informação *a posteriori*		
Parâmetro	Média	Intervalo de credibilidade 95%	
γ_{10}	0,2516	0,119	0,391
γ_{20}	-0,3988	-0,525	-0,274
γ_{30}	2,9812	2,918	3,048
γ_{40}	2,7066	2,648	2,764
γ_{50}	-3,1251	-3,207	-3,040
γ_{11}	-0,0800	-0,268	0,106
γ_{21}	-0,6477	-0,844	-0,450
γ_{31}	0,0122	-0,157	0,167
γ_{41}	0,7115	0,538	0,882
γ_{51}	0,5902	0,426	0,759

Tabela 8 - A Tabela 8 apresenta estimativas Bayesianas de outros parâmetros que podem ser obtidas por meio daqueles estimados pelo modelo proposto. Estimativas com seus respectivos intervalos de credibilidade 95%

Cor da pele	Teste	Parâmetro	Informação *a posteriori*	
			Média	Intervalo de credibilidade 95%
Brancos	PSA	S_1	0,563	0,529 0,597
		E_1	0,952	0,948 0,955
	DRE	S_2	0,402	0,371 0,432
		E_2	0,937	0,933 0,941
		p	0,042	0,038 0,046
		rTVP	1,402	1,288 1,524
		rTFP	0,772	0,710 0,836
Negros	PSA	S_1	0,543	0,486 0,596
		E_1	0,952	0,943 0,959
	DRE	S_2	0,261	0,219 0,306
		E_2	0,968	0,962 0,973
		p	0,074	0,061 0,086
		rTVP	2,097	1,727 2,513
		rTFP	1,515	1,194 1,897

Essas informações *a posteriori* comparam explicitamente as medidas de sensibilidade dos testes e a prevalência do câncer de próstata, entre os indivíduos brancos e negros.

Outra quantidade que pode ser estimada pelo modelo Bayesiano é a razão de prevalências de câncer prostático segundo a cor da pele. Usando as amostras de Gibbs geradas, a estimativa da razão de prevalências é 1,750 (com intervalo de credibilidade 95% dado por (1,496; 2,039)), ou seja, a doença seria aproximadamente 75% mais prevalente nos indivíduos negros.

5 Discussão

Os dados de câncer de próstata introduzidos por Smith et al. (1997) também foram explorados em um artigo de Pepe e Alonzo (2001), no qual foi proposto um modelo de regressão (não Bayesiano) para as taxas de falsos positivos e verdadeiros positivos. Por meio desse modelo, esses autores mostraram que a sensibilidade do PSA é 1,40 vez a sensibilidade do DRE para os homens brancos e 2,11 vezes a sensibilidade do DRE para os negros. O modelo Bayesiano leva a estimativas bastante próximas: a *rTVP* é estimada em 1,402 para os indivíduos brancos e em 2,097 para os negros (Tabela 8). Pepe e Alonzo (2001) estimaram que a *rTFP* para os negros é 1,84 vezes a *rTFP* para os brancos. No modelo aqui utilizado, esta quantidade foi 1,95 (com intervalo de credibilidade 95% dado por (1,565; 2,444)).

É importante salientar que inferências Bayesianas mais precisas poderiam ser obtidas considerando distribuições *a priori* informativas, as quais poderiam ser justificadas a partir de opiniões de especialistas, o que geralmente é comum na área médica.

Dependendo do número de testes diagnósticos sob investigação, os métodos freqüentistas podem apresentar problemas de não identificabilidade na estimação das medidas de desempenho. Esse problema se intensifica quando temos um número de parâmetros desconhecidos superior ao número de variáveis. Por outro lado, os métodos Bayesianos apresentam-se como uma alternativa viável.

MARTINEZ, E. Z.; ACHCAR J. A.; LOUZADA-NETO, F. Bayesian analysis of the performance of two diagnostic tests when negative individuals on both tests are not verified by a gold standard. *Rev. Mat. Estat.*, São Paulo, v.22, n.3, p.21-32, 2004.

- *ABSTRACT: The performance of a diagnostic test is usually summarised by its sensitivity and specificity. Sensitivity is the probability of a positive result, given that the individual is truly diseased, and specificity is the probability of a negative result, given a nondiseased individual. These measures are obtained by comparing the test outcomes and the results of a reference test generically denominated gold standard. However, in many applied problems the gold standard is not available for those individuals with negative results on both tests. In this context, we develop a Bayesian inference procedure for performance measures estimation. We also present an extension of this procedure, involving inclusion of covariates. This Bayesian approach is based on Markov Chain Monte Carlo methods. As an example, we apply the proposed method to a real data set obtained from the medical literature.*

- *KEYWORDS: Sensitivity; specificity; diagnostic tests; Bayesian methods.*

Referências

ALONZO, T. A.; BRAUN, T. M.; MOSKOWITZ C. S. Small sample estimation of relative accuracy for binary screening tests. *Stat. Med.*, Chichester, v.23, n.1, p.21-34, 2004

BEST, N. G.; COWLES, M. K.; VINES, S. K. *CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, version 0.3*. Cambridge: Medical Research Council Biostatistics Unit, 1995. Não paginado.

BOX, G. E. P.; TIAO, G. C. *Bayesian inference in statistical analysis*. Boston: Addison Wesley, 1973. 608p.

CARLIN, B. P.; LOUIS, T. A. Bayes and empirical Bayes methods for data analysis. 2nd ed. London: Chapman & Hall/CRC, 2000. 440p.

CHENG, H.; MACALUSO, M. Comparison of the accuracy of two tests with a confirmatory procedure limited to positive results. *Epidemiology*, Baltimore, v.8, n.1, p.104-106, 1997.

CHENG, H.; MACALUSO, M.; HARDIN, J. M. Validity and coverage of estimates of relative accuracy. *Ann. Epidemiol.*, Raleigh, v.10, p.251-260, 2000.

CHOCK, C. et al. Comparing dichotomous screening tests when individuals negative on both tests are not verified. *J. Clin. Epidemiol.*, New York, v.50, p.1211-1217, 1997.

CLAEYS, P. et al. Performance of the acetic acid test when used in field conditions as a screening test for cervical cancer. *Trop. Med. Int. Health*, Oxford, v.8, n.8, p.704-709, 2003.

CRAWFORD, E. D. Epidemiology of prostate cancer. *Urology*, Belle Mead, v.62, p.3-12, 2003. Supplement.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: BERNARDO, J. M. et al. (Ed.). *Bayesian Statistics 4*. Oxford: Clarendom Press, 1992. p.169-194.

JOSEPH, L.; GYORKOS, T. W.; COUPAL, L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the presence of a gold standard. *Am. J. Epidemiol.*, Baltimore, v.141, n.3, p.263-272, 1995.

PEPE, M. S.; ALONZO, T. A. Comparing disease screening tests when true disease status is ascertained only for screen positives. *Biostatistics*, Oxford, v.2, n.3, p.249-260, 2001.

SAS INSTITUTE. *SAS/IML user's guide, version 8*. Cary: SAS Publishing, 1999. 844p.

SCHATZKIN, A. et al. Comparing new and old screening tests when a reference procedure cannot be performed on all screenees. Example of automated cytometry for early detection of cervical cancer. *Am. J. Epidemiol.*, Baltimore, v.125, n.4, p.672-678, 1987.

SMITH, A. F. M.; ROBERTS, G. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. R. Stat. Soc. B*, Cambridge, v.55, p.3-23, 1993.

SMITH, D. S.; BULLOCK, A. D.; CATALONA, W. J. Racial differences in operating characteristics of prostate cancer screening tests. *J. Urol.*, Baltimore, v.158, n.5, p.1861-1865, 1997.

SOX, H. C. Probability theory in the use of diagnostic test: an introduction to critical study of the literature. *Ann. Intern. Med.*, Philadelphia, v.104, n.1, p.60-66, 1986.

TANNER, M. A.; WONG, W. H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.*, Alexandria, v.82, n.398, p.528-540, 1987.

VAN DER MERWE, L.; MARITZ, S. Estimating the conditional false-positive rate for semi-latent data. *Epidemiology*, Baltimore, v.13, n.4, p.424-430, 2002.

WALTER, S. D. Estimation of test sensitivity and specificity when disease confirmation is limited to positive results. *Epidemiology*, Baltimore, v.10, n.1, p.67-72, 1999.

Recebido em 25.05.2004.

Aprovado após revisão em 16.12.2004.