

Correção da prevalência autorreferida em estudos epidemiológicos com grandes amostras

Correction of self-reported prevalence in epidemiological studies with large samples

Corrección de la prevalencia autorreferida en estudios epidemiológicos con grandes muestras

Jessica Pronestino de Lima Moreira ¹
Renan Moritz Varnier Rodrigues de Almeida ²
Nei Carlos dos Santos Rocha ³
Ronir Raggio Luiz ¹

doi:10.1590/0102-311X00050816

Resumo

As prevalências de doenças são úteis para a formulação e avaliação de políticas públicas. A medida autorreferida é comumente utilizada por ser fácil de ser coletada e não exigir treinamento específico em saúde ou custo adicional. Todavia, esse processo de mensuração pode gerar uma medida enviesada. Objetivou-se apresentar os métodos existentes para ajustar a prevalência, com base na autorreferida, focando nos problemas computacionais no caso de amostras grandes e propondo uma solução alternativa. Os métodos foram divididos em: algébrico, de simples execução, porém não é aplicável em qualquer combinação de prevalência autorreferida, especificidade e sensibilidade; e Bayesiano, que não apresenta a restrição da estratégia anterior, mas apresenta problemas computacionais na sua aplicação em computadores pessoais para amostras grandes. Esses problemas impedem a implantação direta do método já existente, havendo a necessidade da apresentação de uma estratégia aproximada que viabilize a estimação. O método empírico proposto para a aplicação em amostras grandes consiste em reduzir o tamanho da amostra até o limite máximo possível de ser calculado pelo software, mantendo a proporção de doentes. O método foi considerado adequado, pois converge para o verdadeiro valor. No exemplo, uma prevalência autorreferida de 5%, com sensibilidade = 0,4 e especificidade = 0,9 foi corrigida para 0,17% (IC95%: 0,10-0,24). O estudo apresentou os métodos existentes para ajuste de prevalências, bem como uma nova estratégia para prevalências oriundas de grandes amostras, permitindo a obtenção de estimativas mais próximas às verdadeiras, sem a necessidade de mensurar diretamente todos os indivíduos.

Prevalência; Estudos Epidemiológicos; Estudos Transversais

¹ Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

² Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

³ Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

Correspondência

J. P. L. Moreira
Rua Joaquim Rego 29, Rio de Janeiro, RJ 21021-590, Brasil.
jessicapronestino@yahoo.com.br



Introdução

Estimativas de prevalências são, frequentemente, derivadas de inquéritos amostrais de base populacional, que relacionam características de saúde e condições de vida em uma população, sendo fundamentais na formulação e avaliação de políticas públicas^{1,2}. Cada inquérito tem sua metodologia e periodicidade próprias e, dentre os principais, destacam-se o *National Health Interview Survey* (NHIS) e o *National Health and Nutrition Examination Survey* (NHANES), nos Estados Unidos, o *General Health Survey* (GHS) e o *Health Survey for England* (HSE), na Inglaterra¹. No Brasil, entre os grandes inquéritos em saúde, destacam-se o VIGITEL (*Vigilância de Doenças Crônicas por Telefone, o Inquérito Domiciliar sobre Comportamentos de Risco e Morbidade Referida de Doenças e Agravos Não Transmissíveis*), realizado pelo Instituto Nacional de Câncer José de Alencar Gomes da Silva (INCA), e a *Pesquisa Nacional por Amostra de Domicílios* (PNAD) e seus suplementos saúde, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Em 2013, o Suplemento Saúde da PNAD deu lugar à *Pesquisa Nacional de Saúde* (PNS), que, além de entrevistas, como na pesquisa anterior, realiza também exames clínicos, fornecendo informações ainda mais ricas e detalhadas sobre a saúde das populações e deverá manter a periodicidade de 5 anos³.

Não obstante o papel fundamental dos inquéritos para a estimação de prevalência de doenças, estes podem representar altos custos e rotinas complexas de coleta de dados, especialmente quando envolvem diagnósticos de doenças que demandam profissionais e/ou equipamentos especializados. Uma solução alternativa que tem sido empregada é a medida autorreferida da doença, que é fácil de ser coletada em grandes populações e não exige treinamento específico em saúde ou custo adicional. Todavia, esse processo de mensuração pode gerar uma medida de prevalência enviesada, já que, ao serem perguntados, os entrevistados podem se classificar como doentes sem ter a doença – os falso-positivos – e outros que têm a doença podem se classificar como não doentes – os falso-negativos. Assim, se uma prevalência autorreferida é informada e assimilada como equivalente à prevalência real, sem atentar sobre a necessidade de correção, ficamos inseguros em reconhecer a validade da informação.

Por outro lado, se a sensibilidade e a especificidade da pergunta associada a esse procedimento são conhecidas, a prevalência poderia, à primeira vista, ser prontamente corrigida matematicamente e com alguma simplicidade algé-

brica^{4,5,6}. No entanto, surpreendentemente, essa correção pode estar associada a dificuldades matemáticas ou até mesmo computacionais, quando oriunda de amostras grandes. Nesse sentido, o presente artigo tem o objetivo de apresentar os métodos existentes para realizar o ajuste da prevalência corrigida, com base na prevalência autorreferida, bem como as questões metodológicas relacionadas à correção, focando especialmente nos problemas computacionais encontrados no caso de amostras grandes, propondo, ainda, uma solução alternativa.

Métodos

Método algébrico

Não é difícil deduzir que o número de casos autorreferidos de uma doença é igual à soma daqueles que se declaram acertadamente doentes com aqueles que se declaram erroneamente doentes. Sendo a sensibilidade (S) a proporção daqueles verdadeiramente doentes que assim se declararam, e a especificidade (E) a proporção de indivíduos verdadeiramente sem a doença e que se declaram não doentes, é possível escrever, portanto, a seguinte equação:

$$\text{casos aparentes} = (\text{casos verdadeiros}) \times S + (\text{não casos verdadeiros}) \times (1 - E)$$

Buck & Gart⁷, em 1966, Levy & Kass⁸, em 1970, e Rogan & Gladen⁴, em 1978, formalizaram uma estimativa para a correção de medidas obtidas por testes de rastreamento não perfeitos, baseando-se na equação anterior. Assim, relativizando os casos autorreferidos pelo tamanho da amostra (n), encontramos:

$$\frac{\text{casos autorreferidos}}{n} = \frac{(\text{casos verdadeiros})S}{n} + \frac{(\text{não casos verdadeiros})(1 - E)}{n}$$

Logo, sendo: p_a = prevalência autorreferida e p_r = prevalência real, encontramos:

$$p_a = p_r S + (1 - p_r)(1 - E)$$

Nesse sentido, pode-se reescrever a expressão anterior, evidenciando a prevalência real como função da prevalência autorreferida e dos valores de sensibilidade e especificidade, a saber:

$$p_a = p_r S + (1 - p_r)(1 - E)$$

$$p_a = p_r S + (1 - E - p_r + p_r \times E)$$

$$p_a = p_r(S + E - 1) + 1 - E$$

$$p_a + E - 1 = p_r(S + E - 1)$$

Logo:

$$p_r = \frac{p_a + E - 1}{S + E - 1},$$

demonstrando que a prevalência real é uma função linear da prevalência aparente e depende claramente da sensibilidade e da especificidade.

A suposta simplicidade da solução mostrada anteriormente, por outro lado, está associada a algumas condições matemáticas. Para que a expressão retorne valores de prevalências interpretáveis – no intervalo [0,1] – é preciso satisfazer a seguinte desigualdade:

$$0 \leq \frac{p_a + E - 1}{S + E - 1} \leq 1$$

Resolvendo a desigualdade sob a condição $S + E > 1$, encontramos:

$$\begin{aligned} \frac{p_a + E - 1}{S + E - 1} &\leq 1 \\ p_a + E - 1 &\leq S + E - 1 \\ p_a &\leq S \end{aligned}$$

Já com relação à primeira desigualdade, ainda sob a condição $S + E > 1$, encontramos:

$$\begin{aligned} 0 &\leq \frac{p_a + E - 1}{S + E - 1} \\ p_a + E - 1 &\geq 0 \\ p_a &\geq 1 - E \end{aligned}$$

Logo, é preciso que as duas condições sejam satisfeitas concomitantemente para que a prevalência real apresente valores interpretáveis, ou seja, $1 - E \leq p_a \leq S$.

A consequência do método algébrico, pela imposição dessa dupla condição, é que nem todas as prevalências autorreferidas são passíveis de ser corrigidas, mas somente aquelas restritas ao intervalo entre o complementar da especificidade e a sensibilidade. A Figura 1, representando a prevalência real em função da prevalência aparente, ilustra graficamente as consequências das condições anteriores, assumindo, como exemplo, dois pares de sensibilidade e especificidade. Observa-se que com valores de sensibilidade igual a 0,5 e especificidade igual a 0,9, para se obter valores de prevalência real no intervalo esperado [0,1] é necessário que a prevalência autorreferida esteja dentro do intervalo [0,1; 0,5], ou seja, satisfaça à condição anterior $0,1 \leq p_a \leq 0,5$. Já para uma sensibilidade de 0,8 e a mesma especificidade, observa-se a necessidade da prevalência autorreferida estar no intervalo [0,1; 0,8].

Em síntese, caso as condições necessárias não sejam satisfeitas, a prevalência corrigida apresentará valores impossíveis (negativos ou maiores que um). Rogan & Gladen⁴ sugeriram,

como solução, atribuir os valores “zero” quando a prevalência real (p_r) apresentar valores negativos, e “um” quando for maior que um. Não obstante essa solução resolver o problema do ponto de vista prático, ela é claramente insatisfatória e ineficiente, inviabilizando o cálculo do intervalo de confiança das estimativas de prevalências⁵, procedimento fundamental para o conhecimento das incertezas associadas a este tipo de estudo.

Método Bayesiano

A fim de corrigir a limitação do método algébrico que, como visto, pode não apresentar valores de prevalência corrigida no intervalo [0,1], Lew & Levy⁵ propuseram um estimador Bayesiano para p_r (p'_r), baseado na sua distribuição *a posteriori*, obtida por meio de uma distribuição *a priori* não informativa (distribuição uniforme em [0,1]) e pela função de verossimilhança, estimando assim valores para a prevalência autorreferida, de forma a sempre retornar valores de prevalência real no intervalo [0,1].

Além disso, apresenta valores próximos dos valores estimados pelo método algébrico na maioria das situações, já que ambos são estimadores consistentes e coincidem com o estimador de máxima verossimilhança da prevalência real (p_r), quando a sensibilidade e a especificidade do teste são conhecidas⁸. Isso significa que a sua estimativa se aproxima do verdadeiro valor do parâmetro na medida em que aumenta o tamanho da amostra⁹.

O estimador Bayesiano da prevalência é dado por:

$$p'_r = \frac{d + E - 1}{S + E - 1}$$

em que:

$$d = \frac{\int_{1-E}^S p_a^{x+1} (1-p_a)^{n-x} dp_a}{\int_{1-E}^S p_a^x (1-p_a)^{n-x} dp_a}$$

n = tamanho da amostra

x = número de casos prevalentes (doentes)

Para o cálculo do intervalo de confiança da prevalência corrigida foi utilizada a aproximação sugerida por Lew & Levy⁵, dada por:

$$p'_r \pm 1,96 SE(p'_r),$$

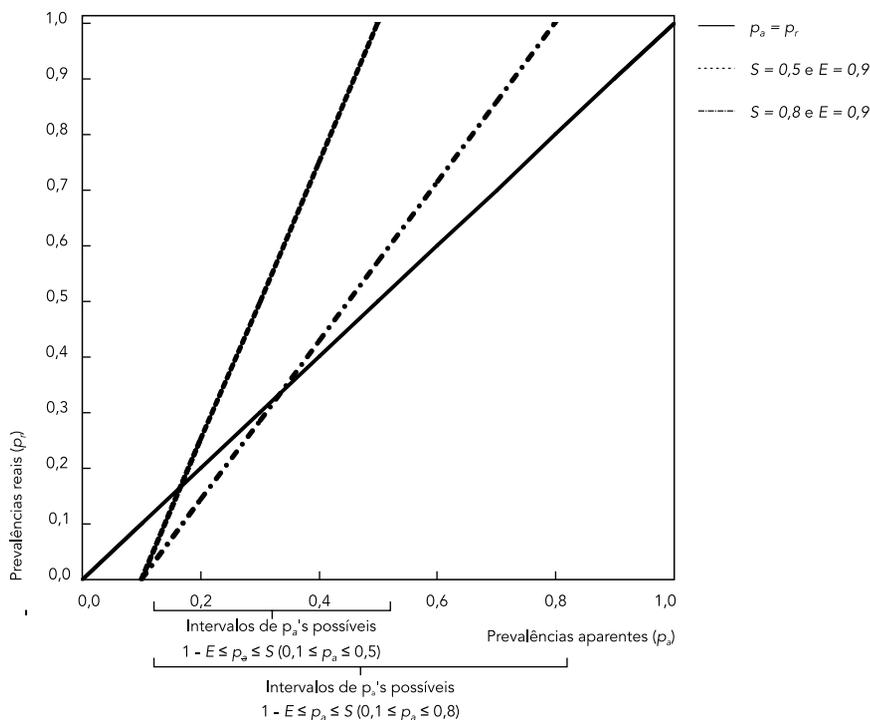
em que:

$$SE(p'_r) = \text{erro-padrão} = \sqrt{\frac{p'_r(1-p'_r)}{[n(S+E-1)]^2}}$$

Portanto, a única diferença deste método para o método algébrico é que, em lugar de se usar

Figura 1

Relação da prevalência real (p_r) com a prevalência autorreferida (p_a) pelo método algébrico.



E: especificidade; S: sensibilidade.

a prevalência autorreferida (p_a), utiliza-se uma função da p_a dada por d , com $1 - E \leq d \leq S$.

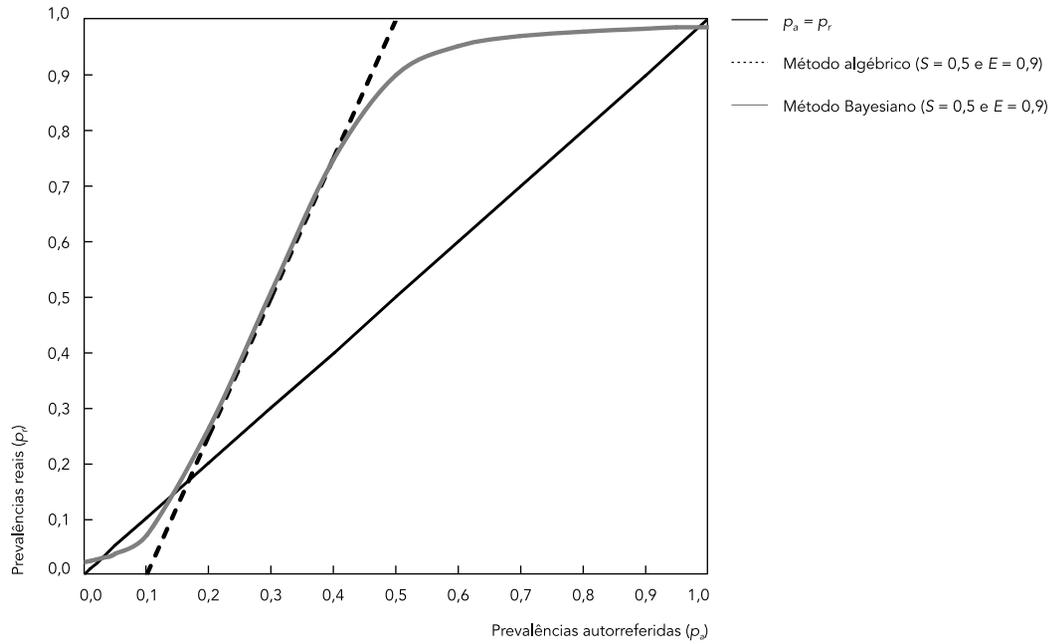
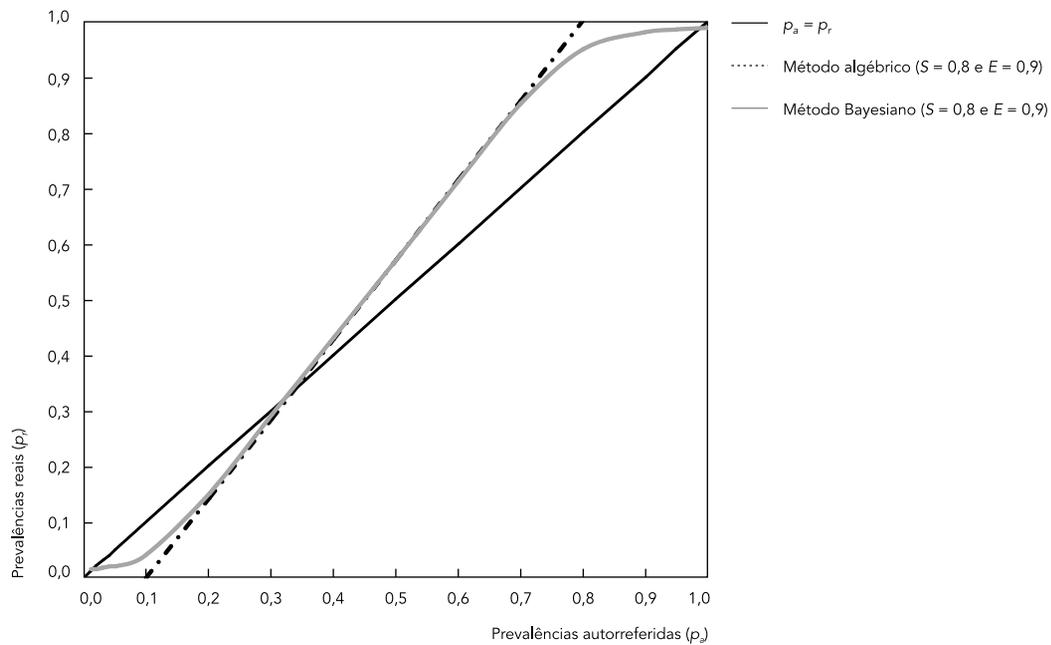
A Figura 2, tomando-se os mesmos dois pares de especificidade e sensibilidade utilizados na Figura 1, ilustra a solução Bayesiana comparativamente à algébrica, evidenciando claramente que a diferença se dá nas extremidades. Fica claro que nas Figuras 2a e 2b, a função se torna sigmoide, o que pode ser facilmente percebido pelas extremidades das figuras, possibilitando a atribuição de valores de prevalência real no intervalo válido para quaisquer valores de prevalência autorreferida. Ou seja, os resultados variam sempre no intervalo $[0,1]$, sem que haja necessidade de restrição de sensibilidade e especificidade, como acontece no método algébrico.

Não obstante a solução analítica do problema, seu cálculo não é imediato e depende de uma integração numérica. Lew & Levy⁵ sugerem, para amostras de tamanho entre 20 e 100, uma aproximação do estimador Bayesiano, para que o valor de d seja obtido rapidamente, sem a necessidade de se utilizarem programas específicos para a integração. Karaagaoglu¹⁰ também apre-

sentou uma proposta alternativa para que não fosse necessário o uso de programa matemático, mas esta substituição só é válida para tamanhos amostrais até 200. Entretanto, acima desse valor é necessário o uso de software. Além disso, em situações concretas de estimação de prevalência são utilizados tamanhos amostrais substancialmente maiores do que esse. Contudo, mesmo em programas específicos, como o Maple ou Matlab, essa integral é de resolução surpreendentemente complexa para amostras grandes. O tamanho máximo tolerável para se calcular em computadores comuns depende da prevalência autorreferida. Amostras entre 1.000 e 4.000 são normalmente bem toleradas em computador pessoal. Já acima desses valores, podem apresentar problemas para efetuar o cálculo. Nesse sentido, a solução pelo método Bayesiano é prejudicada ao utilizar grandes inquéritos de saúde, em que os tamanhos amostrais comumente passam de 100 mil indivíduos, exigindo uma abordagem alternativa aproximada para contornar este problema computacional.

Figura 2

Relação da prevalência real (p_r) com a prevalência autorreferida (p_a) pelos métodos algébrico e Bayesiano.

Figura 2a) Sensibilidade (S) = 0,5 e Especificidade (E) = 0,9Figura 2b) Sensibilidade (S) = 0,8 e Especificidade (E) = 0,9

Ajuste da prevalência em amostras grandes

No contexto de amostras grandes, as soluções apresentadas anteriormente (algébrica e Bayesiana) apresentam problemas matemáticos e/ou computacionais. Uma solução empírica e aproximada possível refere-se à redução do tamanho da amostra. A proposta consiste em reduzir a amostra até o limite possível de ser calculado pelo programa matemático, mantendo a mesma proporção de pessoas doentes encontradas na amostra total, ou seja, mantendo a prevalência autorreferida. Esse valor de d encontrado para o tamanho de amostra máximo reduzido converge satisfatoriamente, de modo a ser uma estimativa boa para o valor de d real desconhecido. A garantia de que essa aproximação converge para o verdadeiro valor de d provém do Teorema da Convergência Dominada ¹¹.

Resultados

Aplicação do ajuste da prevalência em amostra grande

Supondo uma doença com prevalência autorreferida de 5%, estimada com base em uma amostra de 140 mil pessoas, com sensibilidade e especificidade da pergunta iguais a 0,4 e 0,9, respectivamente. Essa prevalência é aproximadamente igual à da hipertensão arterial sistêmica no Brasil, na faixa etária dos adultos jovens, pela PNAD, realizada pelo IBGE, em 2008. Assim, os valores necessários para a correção da prevalência são:

$$E (\text{especificidade}) = 0,9$$

$$S (\text{sensibilidade}) = 0,4$$

$$p_a (\text{prevalência autorreferida}) = 0,05$$

$$n (\text{número de pessoas investigadas}) = 140.000$$

$$x (\text{número de pessoas que se declararam doentes}) = 7.000$$

Verifica-se que os valores antes mostrados não satisfazem à seguinte condição: $1 - E \leq p_a \leq S$, ou seja, a prevalência autorreferida de 5% está fora do intervalo aceitável para correção pelo método algébrico, que é $0,1 \leq p_a \leq 0,4$. Dessa forma, o resultado da correção é negativo, conforme a seguir:

$$p_r = \frac{0,05 + 0,9 - 1}{0,4 + 0,9 - 1}$$

$$p_r = - 0,16$$

Nessa situação, o resultado pela abordagem algébrica seria, portanto, estimado como 0 (zero), o que certamente não é uma boa estratégia para o ajuste da prevalência real.

Utilizando-se o método Bayesiano para a correção da prevalência, encontramos:

$$p_r' = \frac{d + 0,9 - 1}{0,4 + 0,9 - 1},$$

em que d é resultado da seguinte razão de integrais:

$$d = \frac{\int_{1-0,9}^{0,4} p^{7.000+1} (1-p)^{140.000-7.000} dp}{\int_{1-0,9}^{0,4} p^{7.000} (1-p)^{140.000-7.000} dp}$$

Como dito, essa razão poderia ser calculada em programas matemáticos. Entretanto, os expoentes muito grandes nos integrandos de d (com numerador e denominador que se tornam muito pequenos quando n e/ou x são grandes) acarretam um problema computacional, por falta de precisão dos programas matemáticos. Essa falta de precisão faz com que os programas encontrem a razão indefinida matematicamente $0/0$, pois os resultados das integrais do numerador e do denominador são valores reais muito pequenos, inviabilizando, portanto, o cálculo do valor de d e, conseqüentemente, da prevalência real. Alternativamente, a solução empírica proposta de redução do tamanho da amostra pode ser aplicada.

Com o objetivo de validar a convergência numérica proposta, foi feita uma simulação com valores calculados para d em amostras variando de 100 a 3.400 (mantendo a proporção do número de doentes x), sendo este último o tamanho amostral máximo possível de ser calculado em computadores pessoais padrão para esta prevalência, conforme pode ser verificado na Figura 3, contexto este em que o valor exato de d nos é conhecido e igual a 0,1005047065, para o tamanho amostral máximo tolerado de $n = 3.400$ e $x = 170$ (mantendo a prevalência autorreferida de 5%). Uma vez reconhecido que o limite dessa função converge para o verdadeiro valor de d , como garantido pelo Teorema da Convergência Dominada (Figura 4), podemos estimar o limite ou utilizar o valor máximo possível de ser calculado. Nesse caso, tamanhos amostrais a partir de 3 mil já produziram boas estimativas para d , tendo em vista que a diferença nos valores de d para o novo d "real" possível (com $n = 3.400$) passa a ser observada apenas na quarta casa decimal, com um erro relativo mínimo em torno de 0,04%. Portanto, o valor de d calculado para o tamanho amostral máximo possível converge para o verdadeiro valor, mostrando-se uma boa estimativa do d real, não sendo este possível de se acessar com amostras tão grandes.

Figura 3

Simulação dos valores de d , para determinados tamanhos de amostra.

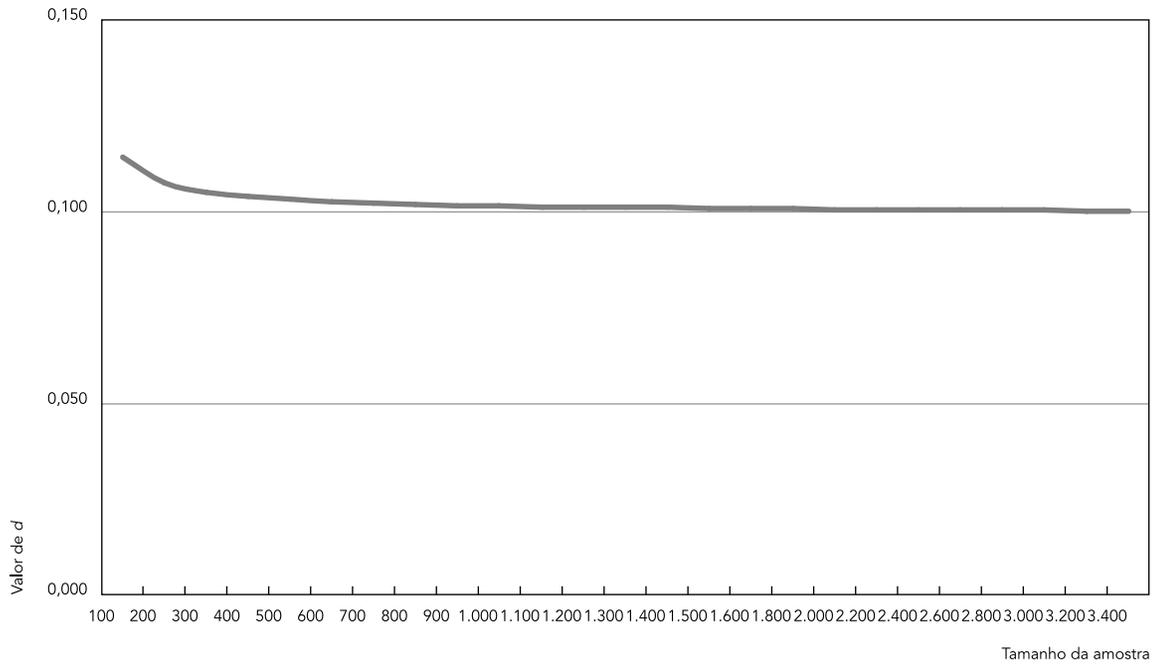


Figura 4

Simulação dos valores de d , para determinados tamanhos de amostra.

A essência do método de aproximação, via Teorema da Convergência Dominada, consiste em construir duas seqüências $(n_k)_{k \geq 1}$ e $(x_k)_{k \geq 1}$ tais que $n_k \rightarrow n$, $x_k \rightarrow x$, de maneira que

$$\frac{\int_{1-E}^S p_a^{x_k+1} (1-p_a)^{n_k-x_k} dp_a}{\int_{1-E}^S p_a^{x_k} (1-p_a)^{n_k-x_k} dp_a} \xrightarrow{k \rightarrow \infty} d = \frac{\int_{1-E}^S p_a^{x+1} (1-p_a)^{n-x} dp_a}{\int_{1-E}^S p_a^x (1-p_a)^{n-x} dp_a}.$$

Ressaltamos uma vez mais a importância de a seqüência atender ao critério $\frac{x_k - x}{n_k} \rightarrow 0$, para todo $k \geq 1$, visto que, em princípio, não saberemos o momento de parada da aproximação que satisfaça ao erro relativo tolerado e, portanto, é necessário resguardar a similaridade da prevalência da doença observada, em qualquer ponto de parada.

Assim, tomando-se esse valor de d e aplicando-o à expressão do estimador Bayesiano, verifica-se que a prevalência ajustada pela sensibilidade e especificidade, com base na prevalência autorreferida é:

$$p'_r = \frac{0,1005047065 + 0,9 - 1}{0,4 + 0,9 - 1} = 0,17\%.$$

Portanto, utilizando-se essa abordagem, a prevalência ajustada dessa doença passa a ser de apenas 0,17%, e não de 5%, como indicado pela prevalência autorreferida e tampouco 0%, segundo a abordagem algébrica de Rogan & Gladen ⁴.

Aplicando a fórmula do erro-padrão, encontramos:

$$\text{var}(p'_r) = \frac{p'_r(1 - p'_r)}{[n(S + E - 1)]} = \frac{0,0017(1 - 0,0017)}{140.000(0,4 + 0,9 - 1)} = 0,000083538129105$$

$$SE(p'_r) = [\text{var}(p'_r)]^{1/2} = 0,000719326$$

Dessa forma, verificamos que o intervalo de confiança estimado da prevalência corrigida é (0,10%; 0,24%).

Discussão

O ajuste da prevalência real de uma doença, baseando-se em seus valores autorreferidos, é de extrema importância no campo da saúde pública, pois permite a estimação da magnitude da doença na população, de forma mais aproximada da real, aprimorando as informações necessárias para tomadas de decisão. No exemplo anterior, uma prevalência aparente de 5% foi corrigida para 0,17%. No entanto, os problemas computacionais encontrados em amostras grandes impedem a implantação dos métodos já existentes, havendo a necessidade da apresentação de uma estratégia aproximada, que viabilize prontamente a estimação.

Estudos que se preocupam em corrigir a prevalência com base no valor autorreferido não são muito frequentes. Em um trabalho realizado na França, para ajuste da prevalência autorreferida em doenças como osteoartrite, verificou-se que havia uma subestimação da prevalência antes da correção, que, no caso de osteoartrite de joelho, passou de 7,9% para 9,1% após a correção¹². Entretanto, o referido trabalho utilizou a correção em duas fases, que consistem em submeter uma subamostra dos indivíduos identificados como positivos e negativos pelo teste de rastreamento, a exames complementares para a confirmação do diagnóstico.

Na mesma linha do presente estudo, Karaagaoglu¹⁰ também propôs uma solução alternativa, que consiste em uma função logística como uma aproximação do estimador Bayesiano, para que não seja necessária a resolução da integral. Apresentou a tabulação dos coeficientes a serem utilizados para o cálculo da verdadeira prevalência para alguns valores de sensibilidade, especificidade e tamanhos amostrais de, no máximo, 200. Apresenta ainda o método de interpolação para outros valores não tabulados, porém não aplicáveis em tamanhos de amostra superiores a 200.

Já Diggle¹³, em 2011, apresentou uma maneira de estimar a prevalência, mesmo quando a sensibilidade e a especificidade do teste forem desconhecidas, permitindo a correção na ausência de estudos prévios sobre o grau de acerto do teste. Entretanto, a proposta de Diggle também apresenta o mesmo problema do método algébrico, podendo resultar de valores de prevalência fora do intervalo interpretável possível (negativos ou acima de 1).

Concluindo, o presente estudo contribui apresentando os métodos existentes, bem como uma nova estratégia para ajuste de prevalências oriundas de grandes inquéritos populacionais, permitindo a obtenção de prevalências próximas às verdadeiras, sem a necessidade de mensurar diretamente todos os indivíduos. Especificamente, pesquisadores que pretendem ajustar a prevalência real de uma doença, baseando-se na medida autorreferida, devem ficar atentos aos problemas causados por amostras grandes e pelas diferenças entre os valores de sensibilidade e especificidade disponíveis.

Colaboradores

Todos os autores participaram em todas as fases do estudo, desde o planejamento até a revisão e aprovação da versão final a ser publicada.

Referências

1. Viacava F. Informações em saúde: a importância dos inquéritos populacionais. *Ciênc Saúde Coletiva* 2002; 7:607-21.
2. Waldman EA, Novaes HMD, Albuquerque MFM, Latorre MRDO, Ribeiro MCSA, Vasconcellos M, et al. Inquéritos populacionais: aspectos metodológicos, operacionais e éticos. *Rev Bras Epidemiol* 2008; 11 Suppl:168-79.
3. Instituto Brasileiro de Geografia e Estatística. Pesquisa Nacional de Saúde 2013: percepção do estado de saúde, estilos de vida e doenças crônicas – Brasil, Grandes Regiões e Unidades da Federação. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística; 2014.
4. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 1978; 107:71-7.
5. Lew RA, Levy PS. Estimation of prevalence on the basis of screening tests. *Stat Med* 1989; 8:1225-30.
6. Klein CH, Costa EA. Os erros de classificação e os resultados de estudos epidemiológicos. *Cad Saúde Pública* 1987; 3:236-49.
7. Buck AA, Gart JJ. Comparison of a screening test and a reference test in epidemiologic studies. I. Indices of agreement and their relation to prevalence. *Am J Epidemiol* 1966; 83:586-92.
8. Levy PS, Kass EH. A three population model for sequential screening for bacteriuria. *Am J Epidemiol* 1970; 91:148-54.
9. Gastwirth JL. The statistical precision of medical screening procedures: application to polygraph and AIDS antibodies test data. *Stat Sci* 1987; 2:213-38.
10. Karaagaoglu E. Estimation of the prevalence of a disease from screening tests. *Turk J Med Sci* 1999; 29:425-30.
11. Neves W, Valle G. Teoria da medida, integração e probabilidade. Rio de Janeiro: Editora UFRJ; 2014.
12. Morvan J, Coste J, Roux CH, Euler-Ziegler L, Sarraux A, Guillemin F. Prevalence in two-phase surveys: accuracy of screening procedure and corrected estimates. *Ann Epidemiol* 2008; 18:261-9.
13. Diggle PJ. Estimating prevalence using an imperfect test. *Epidemiol Res Int* 2011; 2011:5.
14. Aroian LA. Continued fractions for the incomplete beta function. *Annals of Mathematical Statistics* 1941; 6:218-23.

Abstract

Disease prevalence rates are useful when formulating and evaluating public policies. Self-reported measurement is commonly used, since it is easy to collect and does not require specific health training or additional cost. However, this measurement process can produce a biased measure. This study aimed to present the existing methods to adjust prevalence, based on self-report, focusing on computational problems in the case of large samples and proposing an alternative solution. The methods were classified as: algebraic, simple to perform, but not applicable to any combination of self-reported prevalence, specificity, and sensitivity; and Bayesian, which does not have the previous strategy limitations, but displays computational problems when applied to large samples in personal computers. These problems impede the existing method's direct implementation, raising the need to present an approximate strategy to make estimation possible. The empirical method proposed here for application to large samples consists of reducing the sample as far as possible to calculate with the statistical package, maintaining the proportion of patients. We found the method adequate, since it converges with the true value. In the example, a self-reported prevalence of 5% with sensitivity = 0.4 and specificity = 0.9 was corrected to 0.17% (95%CI: 0.10-0.24). The study presented the existing methods for adjusting prevalence rates and a new strategy for prevalence rates in large samples, allowing estimates closer to the true values without the need to directly measure all the individuals.

*Prevalence; Epidemiologic Studies;
Cross-Sectional Studies*

Resumen

Las prevalencias de enfermedades son útiles para la formulación y evaluación de políticas públicas. Las mediciones autorreferidas se utilizan comúnmente por ser fáciles de recoger y no exigir preparación específica en salud o coste adicional. No obstante, este proceso de medición puede generar mediciones sesgadas. Por ello, se presentaron como objetivo de este trabajo los métodos existentes para ajustar la prevalencia, en base a la medición autorreferida, centrándose en problemas computacionales, en el caso de grandes muestras, y proponiendo una solución alternativa. Los métodos se dividieron en: algebraico, de simple ejecución, pese a que no es aplicable en cualquier combinación de prevalencia autorreferida, especificidad y sensibilidad; y Bayesiano, que no presenta una restricción de la estrategia anterior, pero presenta problemas computacionales en su aplicación en ordenadores personales para muestras grandes. Esos problemas impiden la implantación directa del método ya existente, existiendo la necesidad de presentar una estrategia aproximada que viabilice la estimación. El método empírico propuesto para la aplicación en muestras grandes consiste en reducir el tamaño de la muestra hasta el límite máximo posible para que sea calculado por el software, manteniendo la proporción de enfermos. El método fue considerado adecuado, pues converge hacia el verdadero valor. En el ejemplo, una prevalencia autorreferida de un 5%, con una sensibilidad = 0,4 y especificidad = 0,9 fue corregida a 0,17% (IC95%: 0,10-0,24). El estudio presentó los métodos existentes para el ajuste de prevalencias, así como una nueva estrategia para prevalencias procedentes de grandes muestras, permitiendo la obtención de estimativas más próximas a las verdaderas, sin necesidad de realizar mediciones directamente a todos los individuos.

*Prevalencia; Estudios Epidemiológicos;
Estudios Transversales*

Recebido em 28/Mar/2016
Versão final reapresentada em 04/Jul/2016
Aprovado em 24/Ago/2016