Rev Saúde Pública 2011;45(3):617-20

Renan M V R Almeida

Comentários

O papel da plausibilidade na avaliação da pesquisa científica

The role of plausibility in the evaluation of scientific research

RESUMO

O artigo discute o impacto da plausibilidade (probabilidade *a priori*) no resultado de pesquisas científicas, conforme abordagem de Ioannidis, referente ao percentual de hipóteses nulas erroneamente classificadas como "positivas" (estatisticamente significante). A questão "qual fração de resultados positivos é verdadeiramente positiva?", equivalente ao valor preditivo positivo, depende da combinação de hipóteses falsas e positivas em determinada área. Por exemplo, sejam 90% das hipóteses falsas e $\alpha = 0,05$, poder = 0,8: para cada 1.000 hipóteses, 45 (900 x 0,05) serão falso-positivos e 80 (100 x 0,8) verdadeiro-positivos. Assim, a probabilidade de que um resultado positivo seja um falso-positivo é de 45/125. Adicionalmente, o relato de estudos negativos como se fossem positivos contribuiria para a inflação desses valores. Embora essa análise seja de difícil quantificação e provavelmente superestimada, ela tem duas implicações: i) a plausibilidade deve ser considerada na análise da conformidade ética de uma pesquisa e ii) mecanismos de registro de estudo e protocolo devem ser estimulados.

DESCRITORES: Testes de Hipótese. Reprodutibilidade dos Testes. Métodos e Procedimentos Estatísticos.

ABSTRACT

The paper discusses the impact of plausibility (the *a priori* probability) on the results of scientific research, according to the approach proposed by Ioannidis, concerning the percentage of null hypotheses erroneously classified as "positive" (statistically significant). The question "what fraction of positive results are true-positives?", which is equivalent to the positive predictive value, is dependent on the combination of true and false hypotheses within a given area. For example, consider an area in which 90% of hypotheses are false and $\alpha = 0.05$ and power = 0.8: for every 1,000 hypotheses, 45 (900 x 0.05) are false-positives and 80 (100 x 0.8) are true-positives. Therefore, the probability of a positive result being a false-positive is 45/125. In addition, the reporting of negative results as if they were positive would contribute towards an increase in this fraction. Although this analysis is difficult to quantify, and these results are likely be overestimated, it has two implications: i) plausibility should be considered in the analysis of the ethical adequacy of a research proposal, and ii) mechanisms aimed at registering studies and protocols should be encouraged.

DESCRIPTORS: Hypothesis-Testing. Reproducibility of Results. Statistical Methods and Procedures.

Programa de Engenharia Biomédica. Coppe. Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil

Correspondência | Correspondence:
Renan MVR Almeida
Programa de Engenharia Biomédica
Universidade Federal do Rio de Janeiro
Caixa Postal 68510
Cidade Universitária
21945-970 Rio de Janeiro, RJ, Brasil
E-mail: renan.m.v.r.almeida@gmail.com

Recebido: 14/7/2010
Aprovado: 14/11/2010

Artigo disponível em português e inglês em: www.scielo.br/rsp

INTRODUÇÃO

Na metodologia estatística contemporânea,[11] muito utilizada na ciência, normalmente uma hipótese de referência, a "hipótese nula H_0", representa a inexistência de determinado efeito (de onde seu nome). H_0 é considerada "rejeitada" ou "não rejeitada" a partir de uma estatística de teste conveniente (por exemplo, t de Student para a comparação das diferenças entre duas médias). A partir daí, uma estratégia de análise consiste no cálculo de uma probabilidade chamada p-valor, associada a essa estatística. Nos casos em que esse valor encontra-se abaixo do limiar previamente definido (α), diz-se que o efeito existe ou é "estatisticamente significante". Dois tipos de erros são intrínsecos a esse procedimento, chamados Tipo I (rejeitar H_0 sendo ela verdadeira) e Tipo II (não rejeitar H_0 sendo ela falsa). Esses erros ocorrem com probabilidades "α" e "β", respectivamente. Em geral arbitra-se o valor 5% para α e desenhos experimentais comumente desejam um valor de, no máximo, 20% para β (probabilidade de 80% de corretamente rejeitar H_0 falsas – o "poder" do teste).

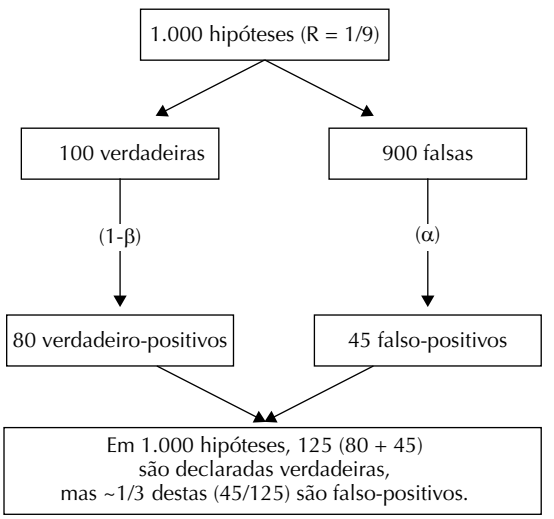
A abordagem tradicional falha ao não considerar o efeito da plausibilidade na avaliação de uma hipótese, entre outros problemas. Particularmente, estatísticos de persuasão não-clássica afirmam que um p-valor pode superestimar a evidência contra uma hipótese, uma vez que o efeito de plausibilidade não é evidente classicamente, em que p = 0,001 é considerado evidência que rejeita igualmente uma hipótese plausível e uma implausível.[3,4]

Assim, o presente artigo discute o impacto da plausibilidade inicial no resultado de pesquisas científicas, conforme abordagem de Ioannidis.[6-9] Essa abordagem refere-se ao percentual de hipóteses nulas H_0 erroneamente classificadas como "resultado positivo" (estatisticamente significante) em vários campos da ciência. Segundo ela, a questão "qual fração de resultados positivos é verdadeiramente positiva?" depende, fundamentalmente, da combinação de hipóteses falsas e positivas testadas em determinada área do conhecimento – a sua probabilidade *a priori*. Essa análise é importante para a compreensão das limitações inerentes à pesquisa científica, particularmente no que se refere aos *priores* (probabilidades iniciais) de um estudo.

ABORDAGEM DE IOANNIDIS

Em recente série de artigos, Ioannidis analisou o papel da replicação e da plausibilidade inicial sobre os resultados da pesquisa científica.[6-9] Seu argumento central foi apresentado em um trabalho com o provocativo título *Why most published research findings are false*,[6] no qual o autor afirma: "*it can be proven that most claimed research findings in most areas of research are false*". Esse trabalho conta com centenas de citações na literatura científica.

Ioannidis sistematizou observações apontadas por outros autores, como Browner & Newman[1] e Sterne & Smith.[15] Assim, os conceitos de erro Tipo I e Tipo II foram apresentados em forma conceitualmente equivalente, de tal maneira que a probabilidade do erro Tipo I foi definida como o percentual de todas as hipóteses H_0, em determinado campo de pesquisa, erroneamente classificadas como estatisticamente significantes; e o erro Tipo II como o percentual das H_0 falsas erroneamente classificadas como estatisticamente não significantes. Dada a observação de um resultado positivo (ou seja, H_0 rejeitada), a probabilidade de que H_0 realmente seja falsa é condicional à fração inicial de hipóteses realmente verdadeiras e realmente falsas testadas. Essa afirmação, análoga ao conceito de valor preditivo positivo, muito usado em testes diagnósticos,[1,2] pode ser compreendida considerando-se os exemplos: i) todas as hipóteses testadas em determinada área são, na realidade, falsas. Nesse caso, 100% dos resultados positivos seriam, na verdade, falsos. ii) 100% das hipóteses são verdadeiras. Analogamente, todos os resultados positivos seriam verdadeiros. iii) em uma área em que 90% das hipóteses testadas sejam falsas (e sejam mantidos os valores convencionais α = 0,05 e poder = 0,8), para cada 1.000 hipóteses, 45 serão falso-positivos (900 x 0,05) e 80 verdadeiro-positivos (100 x 0,8). Assim, dado um resultado positivo, a probabilidade de que seja um falso-positivo é de aproximadamente um terço (45/125) (Figura).



R = proporção hipóteses verdadeiras/hipóteses falsas; 1-β: poder estatístico; α: probabilidade de erro Tipo I. Essa proporção é dependente da combinação inicial das hipóteses verdadeiras e falsas sendo testadas. Figura adaptada de Tabbarok A. Why most published research findings are false. [citado 2010 jul] Disponível em: http://marginalrevolution.com/marginalrevolution/2005/09/why_most_publis.html

Figura. Proporção de hipóteses falsas relativamente ao total de resultados estatisticamente significantes.

Entre os extremos representados pelos casos i e ii, a relação R = H_0 *verdadeiras*/H_0 *falsas* altera o equivalente ao valor preditivo positivo para determinada área do conhecimento (dado um resultado positivo, quanto maior o valor de R, maior a probabilidade de que se trate de um verdadeiro-positivo). Em outras palavras, quanto menor a plausibilidade de um estudo, maior a probabilidade de que um resultado positivo não seja verdadeiro. Esse fenômeno, segundo Ioannidis, auxiliaria a explicar por que, mesmo em revistas científicas de alto impacto, é comum a publicação de estudos contraditórios e não-replicáveis.[6]

Ioannidis introduziu ainda o conceito do "viés u", definido como a probabilidade de um resultado negativo ser relatado erroneamente como positivo por meio do uso seletivo de desfechos secundários, da alteração de pontos de corte, de métodos estatísticos inapropriados ou de fraude. A partir desses conceitos, uma simulação de valores de R e u para diferentes tipos de estudos levou-o à conclusão de que "No contexto descrito, um valor preditivo positivo acima de 50% é bem difícil de ser obtido",[6] (p. 699) e de que "mesmo estudos epidemiológicos com poder adequado podem ter apenas uma chance em cinco de serem verdadeiros, se R = 1:10"[6] (p. 699) (traduções do presente autor). Isso justificaria a afirmação de Ioannidis sobre a prevalência de resultados não-replicáveis em ciência.

ANÁLISE E AVALIAÇÃO

A análise de Ioannidis é contingente em dois aspectos fundamentais: i) que o número de hipóteses falsas em qualquer área do conhecimento é muito maior do que o de hipóteses verdadeiras; e ii) que a suposição de alta taxa u é real (Ioannidis supôs valores u entre 10% e 80%). A primeira pode ser justificada pela própria natureza ousada do trabalho científico, assim como pela pressão constante por resultados, mesmo em campos estéreis ou de progresso lento, mas é de difícil extrapolação para a maioria das áreas de pesquisa. Em contrapartida, na ausência do bem conhecido "efeito gaveta",[13,14] a influência do fenômeno discutido por Ioannidis seria reduzida, uma vez que, se o padrão de R fosse conhecido, o julgamento da verdadeira importância de um resultado positivo seria, em princípio, possível. Mas, de acordo com o discutido, a única forma de avaliar R seria se os resultados negativos e os positivos em uma área de pesquisa fossem conhecidos (por exemplo, se 100 hipóteses sobre determinado fenômeno fossem testadas e 95 fossem reconhecidas como negativas, haveria compatibilidade dos resultados globais com um modelo que supusesse a inexistência do fenômeno). Quanto ao parâmetro u, Goodman & Greenland[5,a] apontam que: i) a definição de u foi enganosa, pois igualava o relato seletivo de desfechos secundários e a fraude direta; e ii) os valores de u assumidos por Ioannidis (10% a 80%) eram especulativos e "dominavam" a simulação, i.e., suas conclusões eram dependentes de níveis elevados de "fraude". Considerando-se que as suposições de Ioannidis são de difícil quantificação, não é possível afirmar que seu efeito refira-se a fração substancial dos resultados científicos.

Segundo os autores, a análise de Ioannidis não distinguiu entre níveis de evidência (em termos de valor-p) contra H_0.[5,8] Ele utilizou dicotomização de resultados como "estatisticamente significantes" e "não significantes" a partir do valor $\alpha = 0,05$ classicamente usado. Entretanto, na prática, é muito incomum que tal dicotomização seja utilizada na apresentação de resultados, sendo preferida a indicação de valores-p específicos.

Irrespectivamente às críticas de Goodman & Greenland[5,a] (que concordam com os pontos centrais da análise de Ioannidis), o efeito discutido é altamente dependente das características específicas de cada área de pesquisa. Ioannidis sugere dois campos como críticos: a pesquisa genômica e a busca por associações entre nutrientes e desfechos epidemiológicos, nos quais as hipóteses são muitas vezes testadas em abordagem heurística e os efeitos são pequenos e de avaliação complexa. Outro exemplo importante[b] é o da área conhecida por Medicina Complementar e Alternativa (MCA), uma vez que não é difícil concluir, a despeito de algumas tentativas,[10] que a única forma de identificar um fio condutor, um núcleo nas inúmeras tendências que se reivindicam MCA, é a clara implausibilidade de suas afirmações. Esse argumento seria conjugado a outros apontados na literatura (como a falta de impacto dos resultados negativos nessa área, a má legitimidade concedida a idéias implausíveis e a alocação inadequada de recursos escassos), indicando a baixa justificativa da realização de estudos de MCA em seres humanos.[12]

Por outro lado, é impossível provar a inexistência de um efeito, pois ele pode encontrar-se abaixo de um limiar de detecção. Além disso, recursos de pesquisa são infinitamente menores do que os necessários para a análise de todos os fenômenos que podem ser propostos. Portanto, pelo menos no âmbito da pesquisa em seres humanos, fenômenos só devem ser investigados se forem relevantes e plausíveis.

CONSIDERAÇÕES FINAIS

Problemas inerentes aos procedimentos da ciência contemporânea facilitam a publicação indevida de resultados aparentemente positivos. Esses problemas relacionam-se à própria plausibilidade dos estudos em determinada área, conectando-se também ao chamado

[a] Goodman S, Greenland S. Assessing the unreliability of the medical literature: a response to "Why most published research findings are false". Baltimore: Johns Hopkins University; 2007[citado 2010 jul]. (Working paper, 135). Disponível em: http://www.bepress.com/jhubiostat/paper135
[b] Novella S. Are most medical studies wrong? *Neurologica Blog*. 2007[citado 2010 jul]. Disponível em: http://theness.com/neurologicablog/?p=8

"efeito gaveta". De acordo com a discussão apresentada, a quantificação da magnitude desses efeitos é difícil, por serem eles, também, dependentes das condições de pesquisa em áreas específicas.

No entanto, destacam-se duas implicações: a primeira diz respeito à importância do princípio operacional reconhecido pela declaração de Helsinque, que em seu artigo 11 declara: "A pesquisa médica envolvendo sujeitos humanos deve estar em conformidade com os princípios científicos geralmente aceitos, deve ser baseada em um amplo conhecimento da literatura científica, em outras fontes relevantes de informação e em experimentação laboratorial e animal apropriada."[16] (tradução do presente autor). A falta de plausibilidade deve ser elevada à condição de violação importante da ética de pesquisa. A segunda refere-se à necessidade do desenvolvimento de mecanismos de registro de estudos[13] que facilitem a detecção e minimização tanto do efeito gaveta quanto de desvios e alterações de protocolo. Esses mecanismos de registro teriam a vantagem de auxiliar a identificação de estudos duplicados e de facilitar a realização de meta-análises, contribuindo, portanto, para a maior transparência e eficiência da pesquisa científica.

REFERÊNCIAS

1. Browner W, Newman TB. Are all significant p values created equal? The analogy between diagnostic tests and clinical research. *JAMA*. 1987;257(18):2459-63.
2. Dawson B, Trapp RG. Basic & Clinical Biostatistics. New York: McGraw-Hill; 2004.
3. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med*. 1999;130(12):995-1004.
4. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med*. 1999;130(12):1005-13.
5. Goodman S, Greenland S. Why most published research findings are false: problems in the analysis. *PLoS Med*. 2007;4(4):e168. DOI:10.1371/journal.pmed.0040168
6. Ioannidis JPA. Why most published research findings are false *PLoS Med*. 2005;2(8):e124. DOI:10.1371/journal.pmed.0020124
7. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research *JAMA*. 2005;294(2):218-28.
8. Ioannidis JPA. Why most published research findings are false: author's reply to Goodman and Greenland. *PLoS Med*. 2007;4(6):e215. DOI:10.1371/journal.pmed.0040215
9. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008;16(4) 640-8. DOI:10.1097/EDE.0b013e31818131e7
10. Manzini T, Martinez EZ, Carvalho ACD. Conhecimento, crença e uso de medicina alternativa e complementar por fonoaudiólogas. *Rev Bras Epidemiol*. 2008;11(2):304-14. DOI:10.1590/S1415-790X2008000200012
11. Moore DS. Estatística Básica e sua Prática. Rio de Janeiro: LTC Editora; 2005.
12. Renkens CNM. Some complementary and alternative therapies are too implausible to be investigated. *Focus Alternat Complement Ther*. 2003;8(3):307-8. Disponível em:
13. Yamey G. Scientists who do not publish trial results are "unethical". *BMJ*. 1999; 319(7215):939.
14. Young NS, Ioannidis JPA, Al-Ubaydli O. Why current publication practices may distort science. *PLoS Med*. 2008;5(10):e201. DOI:10.1371/journal.pmed.0050201
15. Sterne JAC, Smith GD. Sifting the evidence: what is wrong with significance tests? *BMJ*. 2001;322:226-31.
16. World Medical Association. Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects - 2008 version. Ferney-Voltaire; 2008[citado 2010 jul]. Disponível em: http://www.wma.net/en/30publications/10policies/b3/index.html

Trabalho apresentado no VIII Congresso Brasileiro de Bioética, em Búzios, RJ, 2009.
O autor declara não haver conflitos de interesse.