MANIFESTO DE LEIDEN SOBRE MÉTRICAS DE PESQUISA

Diana Hicks[a], Paul Wouters[b], Ludo Waltman[b], Sarah de Rijcke[c] e Ismael Rafols[c, d, e]

a. School of Public Policy, Georgia Institute of Technology, Atlanta, EUA

b. Centre for Science and Technology Studies (CWTS), Universidade de Leiden, Holanda

b. Ingenio (CSIC-UPV), Universidade Politécnica de Valência, Valência, Espanha

c. Science Policy Research Unit (SPRU), Universidade de Sussex, Brighton, Reino Unido

d. Observatoire des Science et des Téchniques (OST-HCERES), Paris, França

(Tradução em Português brasileiro de Hicks et al. The Leiden Manifesto for research metrics. *Nature*, v. 520, p. 429-431, 2015. http://www.sibi.usp.br/programas/bibliometria-e-indicadores-cientificos/manifesto-leiden/)

Cada vez mais se utilizam dados para gerenciar a ciência. As avaliações da pesquisa, que já foram individualizadas, solicitadas e realizadas por pares, atualmente são rotineiras e baseadas em métricas (1). A questão é que agora a avaliação é majoritariamente dependente de dados, ao invés de juízos de valor. As métricas proliferaram: em geral bem intencionadas, nem sempre bem informadas, e frequentemente mal aplicadas. Corremos o risco de prejudicar o sistema da ciência com as próprias ferramentas projetadas para melhorá-lo, uma vez que a avaliação é cada vez mais realizada por instituições sem o devido conhecimento sobre as boas práticas e sobre a interpretação adequada de indicadores.

Anteriormente a 2000, os especialistas utilizavam em suas análises o *Science Citation Index* (SCI) do *Institute for Scientific Information* (ISI), em sua versão em CD-ROM. Em 2002, a Thomson Reuters lançou uma plataforma web integrada, tornando a base *Web of Science* (WoS) acessível a um público mais amplo. Logo surgiram índices de citações concorrentes: a base Scopus, da Elsevier (lançada em 2004) e o Google Scholar (versão beta lançada em 2004). Outras ferramentas baseadas na web surgiram para facilitar a comparação da produtividade da pesquisa institucional e seu impacto, como o InCites (que usa dados da WoS) e o SciVal (com dados da Scopus), bem como aplicativos para analisar perfis individuais de citação com dados do Google Scholar (*Publish or Perish*, lançado em 2007).

Em 2005 Jorge Hirsch, físico da Universidade da Califórnia em San Diego, propôs o Índice h, popularizando a contagem de citações de pesquisadores individuais. O interesse pelo Fator de Impacto de revistas cresceu de forma constante a partir de 1995.

Mais recentemente, ganham impulso métricas relacionadas ao uso social e conversações online – como o F1000 Prime, criado em 2002; o Mendeley, em 2008; e o Altmetric.com, em 2011.

Como cientometristas, cientistas sociais e gestores de pesquisa, temos observado com crescente apreensão a má aplicação generalizada de indicadores na avaliação do desempenho científico. Os exemplos a seguir são apenas alguns de inúmeros casos. Em todo o mundo, as universidades tornaram-se obcecadas com a sua posição nos rankings mundiais (a exemplo do Ranking de Xangai e da lista do *Times Higher Education* - THE), apesar dessas listas serem baseadas, no nosso ponto de vista, em dados imprecisos e indicadores arbitrários.

Algumas instituições solicitam o valor do Índice h dos pesquisadores candidatos a seus postos. Várias decisões de promoção e fomento de universidades baseiam-se nos valores do Índice h e no número de artigos publicados em revistas de "alto impacto". Os currículos dos pesquisadores transformaram-se em espaços para alardear essas pontuações, principalmente na área da Biomedicina. Em todos os lugares, orientadores pressionam prematuramente seus alunos de doutorado a publicar em revistas de "alto impacto" e obter financiamento externo.

Na Escandinávia e na China, algumas universidades distribuem fundos ou bônus para as pesquisas com base em números: por exemplo, por meio do cálculo das pontuações de impacto para alocar recursos baseados no “desempenho individual'', ou concedendo bônus aos pesquisadores para que publiquem em periódicos com Fator de Impacto maior de 15 (2).

Em muitos casos, os pesquisadores e avaliadores ainda exercem um julgamento equilibrado. No entanto, o abuso de métricas da pesquisa tornou-se disseminado demais para ser ignorado.

Assim, apresentamos o Manifesto de Leiden, nomeado após a conferência em que se consolidou (ver http://sti2014.cwts.nl). Seus dez princípios não são novidade para os cientometristas, embora nenhum de nós seria capaz de recitá-los na íntegra, devido à falta de uma codificação integradora até o momento. Luminares do campo da Cientometria, como Eugene Garfield (fundador do ISI), já se referiram a alguns desses princípios (3, 4). Mas esses especialistas não estão presentes quando os avaliadores se reportam aos gestores universitários que também não são especialistas na metodologia pertinente. Os cientistas que procuram a literatura para contestar ou questionar as avaliações só encontram as informações de que necessitam no que são, para eles, periódicos obscuros e de difícil acesso.

Assim, oferecemos essa síntese das melhores práticas de avaliação da pesquisa baseada em métricas, para que os pesquisadores possam confiar em seus avaliadores, e para que os avaliadores possam confiar em seus indicadores.

OS DEZ PRINCÍPIOS

1. A avaliação quantitativa deve dar suporte à avaliação qualitativa especializada.

Os indicadores quantitativos podem corrigir tendências enviesadas da avaliação por pares e facilitar a deliberação. Nesse sentido, devem fortalecer a revisão por pares já emitir julgamentos sobre colegas é difícil sem uma série de informações relevantes. No entanto, os avaliadores não devem ceder à tentação de basear suas decisões apenas em números. Os indicadores não devem substituir o juízo informado. Os tomadores de decisão têm plena responsabilidade por suas avaliações.

2. Medir o desempenho de acordo com a missão da instituição, do grupo ou do pesquisador.

Os objetivos de um programa de pesquisa devem ser indicados no início, e os indicadores utilizados para avaliar seu desempenho devem estar claramente vinculados a esses objetivos. A escolha dos indicadores e de como eles são utilizados deve levar em conta o contexto socioeconômico e cultural mais amplo. Os cientistas tem diversas missões de pesquisa. A pesquisa que avança as fronteiras do conhecimento acadêmico difere da pesquisa que é focada em proporcionar soluções para os problemas da sociedade. A avaliação pode ser baseada em méritos relevantes para as políticas públicas, para a indústria ou para os cidadãos em geral, em vez de méritos baseados em noções acadêmicas de excelência. Não existe um modelo único de avaliação que se aplique a todos os contextos.

3. Proteger a excelência da pesquisa localmente relevante.

Em muitas partes do mundo, a excelência da pesquisa é associada à publicação no idioma Inglês. A lei espanhola, por exemplo, menciona explicitamente a conveniência de que os pesquisadores espanhóis publiquem em revistas de alto impacto. O Fator de Impacto é calculado na Web of Science, que indexa principalmente os periódicos com base nos Estados Unidos e em língua inglesa.

Este viés é particularmente problemático para as Ciências Sociais e Humanidades, áreas mais orientadas para a pesquisa de temas regionais e nacionais. Muitas outras áreas possuem uma dimensão nacional ou regional – a exemplo da Epidemiologia do HIV na África subsaariana.

Este pluralismo e a relevância para a sociedade tendem a ser suprimidos quando se criam artigos de interesse para os guardiões do alto impacto: as revistas em Inglês. Os sociólogos espanhóis altamente citados na *Web of Science* têm trabalhado com modelos abstratos ou com dados dos Estados Unidos. Neste processo, perde-se a especificidade dos sociólogos em revistas espanholas de alto impacto: temas como leis trabalhistas locais, serviços de saúde familiar para idosos ou empregabilidade de imigrantes (5). Os indicadores baseados nas revistas de alta qualidade publicadas em outros idiomas diferentes do Inglês devem identificar e premiar as áreas de pesquisa de interesse local.

4. Manter a coleta de dados e os processos analíticos abertos, transparentes e simples.

A construção das bases de dados necessárias para a avaliação deve observar regras claramente definidas e fixadas antes da conclusão da pesquisa. Esta era a prática comum entre os grupos acadêmicos e comerciais que desenvolveram metodologias de avaliação bibliométrica ao longo de muitas décadas. Tais grupos referenciaram protocolos publicados na literatura revisada por pares. Esta transparência possibilitou o escrutínio das metodologias. Por exemplo, em 2010, o debate público sobre as propriedades técnicas de um importante indicador utilizado por um dos nossos grupos (o Centro de Estudos de Ciência e Tecnologia - *Centre for Science and Technology Studies*, CWTS, da Universidade de Leiden, na Holanda) levou a uma revisão no cálculo deste indicador (6). Os novos operadores do setor privado devem seguir os mesmos padrões; ninguém deve aceitar avaliações saídas de uma caixa-preta.

A simplicidade é uma virtude em um indicador, pois favorece a transparência. Mas métricas simplistas podem promover distorções (ver princípio 7). Os avaliadores devem se esforçar para encontrar o equilíbrio com base em indicadores simples que espelhem com exatidão a complexidade do processo de investigação.

5. Permitir que os avaliados verifiquem os dados e as análises.

Para garantir a qualidade dos dados, todos os pesquisadores incluídos em estudos bibliométricos deveriam poder verificar se suas produções foram corretamente identificadas. Todos os que dirigem e administram os processos de avaliação devem assegurar a precisão dos dados, através de verificação própria ou auditoria de terceiros. As universidades poderiam implementar esse princípio em seus sistemas de informação sobre a pesquisa, o que deveria ser um princípio norteador na seleção de fornecedores desses sistemas. A coleta e processamento de dados precisos e de alta qualidade demandam tempo e dinheiro e devem ser considerados no orçamento institucional.

6. Considerar as diferenças entre áreas nas práticas de publicação e citação.

A melhor prática de avaliação é selecionar um conjunto de possíveis indicadores e permitir que as distintas áreas escolham aqueles que lhes são mais adequados. Há alguns anos, um grupo europeu de historiadores recebeu uma classificação relativamente baixa em uma avaliação nacional por pares, porque escreviam livros em vez de artigos em revistas indexadas na WoS. Estes historiadores tiveram o azar de fazer parte de um departamento de Psicologia. Historiadores e cientistas sociais precisam que os livros e a literatura publicada no idioma nacional sejam incluídos na contagem de publicações; já os cientistas da computação esperam que seus trabalhos apresentados em eventos e conferências sejam levados em conta.

Os valores de citações variam por área: as revistas melhor avaliadas em Matemática têm Fator de Impacto por volta de 3; já as revistas melhor avaliadas em Biologia Celular tem Fator de Impacto em torno de 30. Portanto, é necessário o uso de indicadores normalizados, e o método de normalização mais confiável é baseado em percentuais: cada artigo é ponderado segundo o percentual a que pertence na distribuição de citações em sua área (os melhores 1%, 10% ou 20%, por exemplo). Uma única publicação altamente citada melhora ligeiramente a posição de uma universidade em um ranking baseado em indicadores percentuais, mas pode impulsionar a universidade de uma posição mediana para as primeiras posições em um ranking baseado em médias de citação (7).

7. Basear a avaliação de pesquisadores individuais no juízo qualitativo da sua carreira.

Quanto mais idade você tem, maior será o seu Índice h, mesmo que não publique novos artigos. O Índice h varia por área: os pesquisadores das Ciências da Vida chegam ao topo com 200; os físicos com 100 e cientistas sociais com 20 a 30 (8). Depende da base de dados: há pesquisadores em Ciência da Computação que têm um Índice h de cerca de 10 na WoS, mas de 20 a 30 no Google Scholar (9). Ler e julgar o trabalho de um pesquisador é muito mais adequado do que depender de um número. Mesmo quando se compara um grande número

de pesquisadores, uma abordagem que considere informações diversas sobre o conhecimento, experiência, atividades e influência de cada indivíduo é a melhor.

8. Evite solidez mal colocada e falsa precisão.

Indicadores de ciência e tecnologia são propensos à ambiguidade conceitual e à incerteza, e demandam fortes suposições que não são universalmente aceitas. O significado das contagens de citações, por exemplo, tem sido amplamente discutido. Assim, a melhor prática de avaliação utiliza indicadores múltiplos para fornecer uma imagem mais robusta e plural da pesquisa. Se as incertezas e os erros podem ser quantificados, esta informação deve acompanhar os valores dos indicadores publicados, usando barras de erro, por exemplo. Se isso não for possível, os produtores de indicadores deveriam, pelo menos, evitar a falsa precisão. Por exemplo, o Fator de Impacto de revistas é publicado com três casas decimais para evitar empates. No entanto, dada a ambiguidade conceitual e a variabilidade aleatória das contagens de citações, não faz sentido distinguir as revistas com base em diferenças mínimas do Fator de Impacto. Evite a falsa precisão: apenas uma casa decimal se justifica.

9. Reconhecer os efeitos sistêmicos da avaliação e dos indicadores.

Os indicadores mudam o sistema da pesquisa por meio dos incentivos que estabelecem. Estes efeitos devem ser previstos. Isto significa que um conjunto de indicadores é sempre preferível - um único indicador convida a burlas ou a desvios do objetivo (em que a medida se torna um fim em si). Por exemplo, na década de 1990, a Austrália financiou a pesquisa universitária através de uma fórmula baseada sobretudo no número de artigos publicados pelas instituições. As universidades poderiam calcular o "valor" de um artigo em uma revista revisada por pares; em 2000, o valor era de estimados AUS$ 800 (em torno de US$ 480) para o financiamento da pesquisa. Previsivelmente, o número de artigos publicados por pesquisadores australianos subiu, mas em revistas menos citadas, sugerindo uma queda na qualidade dos artigos (10).

10. Examinar e atualizar os indicadores regularmente.

A missão da pesquisa e os objetivos da avaliação mudam, e o próprio sistema de pesquisa evolui junto. Medidas que anteriormente eram úteis se tornam inadequadas e surgem novos indicadores. Os sistemas de indicadores têm de ser revistos e talvez modificados. Percebendo os efeitos de sua fórmula simplista, em 2010 a Austrália introduziu a iniciativa “Excelência na Pesquisa para a Austrália” (*Excellence in Research for Australia*), mais complexa e com ênfase na qualidade.

Próximos passos

Respeitando esses dez princípios, a avaliação da pesquisa pode desempenhar um papel importante no desenvolvimento da ciência e de suas interações com a sociedade. As métricas da pesquisa podem fornecer informações cruciais que seriam difíceis de reunir ou entender por especialistas individuais. Mas não se deve permitir que essa informação quantitativa se transforme de instrumento em um fim em si.

As melhores decisões são tomadas através da combinação de estatísticas robustas com sensibilidade para a finalidade e a natureza da pesquisa que é avaliada. Tanto a evidência quantitativa quanto a qualitativa são necessárias; cada uma é objetiva à sua maneira. A tomada de decisão na ciência deve ser baseada em processos de alta qualidade informados por dados da mais alta qualidade.

Referências

1. WOUTERS, P. The citation: From culture to infrastructure. In: CRONIN, B.; SUGIMOTO, C. (Eds.). **Beyond Bibliometrics**: Harnessing Multidimensional Indicators of Scholarly Impact. Cambridge, MA: MIT Press, 2014. p. 47–66.

2. SHAO, J.; SHEN, H. The outflow of academic papers from China: why is it happening and can it be stemmed? **Learned Publishing**, v. 24, p. 95–97, 2011.

3. SEGLEN, P. O. Why the impact factor of journals should not be used for evaluating research. **British Medical Journal**, v. 314, n. 7079, p. 498–502, 1997.

4. Garfield, E. J. The history and meaning of the journal impact factor. **Journal of the American Medical Association**, v. 95, n. 1, p. 90–93, 2006.

5. LÓPEZ PIÑEIRO, C.; HICKS, D. Reception of Spanish sociology by domestic and foreign audiences differs and has consequences for evaluation. **Research Evaluation**, v. 24, n. 1, p. 78–89, 2014.

6. VAN RAAN, A. F. J.; VAN LEEUWEN, T. N.; VISSER, M. S. et al. Rivals for the crown: Reply to Opthof and Leydesdorff. **Journal of Informetrics**, v. 4, n. 3, p. 431–435, 2010.

7. WALTMAN, L.; CALERO-MEDINA, C.; KOSTEN, J. et al. The Leiden Ranking 2011/2012: Data Collection, Indicators, and Interpretation. **Journal of the American Society for Information Science and Technology**, v. 63, n. 12, p. 2419-2432, 2012.

8. HIRSCH, J. E. An index to quantify an individual's scientific research output. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 46, p. 16569-16572, 2005.

9. BAR-ILAN, J. Which h-index?—A comparison of WoS, Scopus and Google Scholar. **Scientometrics**, v. 74, n. 2, p. 257–271, 2008.

10. BUTLER, L. Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. **Research Policy**, v. 32, p. 143–155, 2003.