

# A IMPORTÂNCIA DA INTEGRAÇÃO DE DADOS: a produção científica nacional em Zika

Maria Simone de Menezes Alencar

Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, RJ, Brasil  
simone.alencar@unirio.br

Rosany Bochner

Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brasil  
rosany.bochner@icict.fiocruz.br

Daniel Giacometti Amaral

NanoBusiness Informação e Inovação Ltda, Duque de Caxias, RJ, Brasil  
dgiacometti.amaral@gmail.com

## 1 INTRODUÇÃO

Os estudos para avaliação da produção científica baseados em bibliometria usualmente utilizam apenas uma base de dados de artigos científicos, de acordo com a disponibilidade de metadados que a fonte possui e o objetivo do estudo.

Um exemplo disso pode ser observado nos estudos nacionais recentes acerca da produção científica sobre Zika, quando pesquisadores decidiram adotar uma única base para realizar suas buscas, sendo que essa não foi a mesma para todos os trabalhos. Dentre os seis trabalhos analisados, três bases foram apontadas, *Scopus*, *Web of Science (WoS)* e *PubMed*. Araújo, Silva e Guimarães (2016), Ribeiro et al. (2016) e Araújo et al. (2017) adotaram a base *Scopus*; Gheno et al. (2016, 2017) optaram pelo uso da *WoS*; Nunes et al. (2016) utilizaram a base de dados do *PubMed*. Analisando os resultados desses trabalhos foi possível verificar que a base *Scopus* se mostrou mais abrangente do que a *WoS*. No entanto, sabe-se que há trabalhos indexados na *WoS* que não estão na *Scopus* e vice-versa, assim como há trabalhos na *PubMed*, ou mesmo na SciELO, que não estão indexados nas outras bases.



Refletindo sobre esses achados, e na busca por uma pesquisa mais abrangente de produção científica, o objetivo deste estudo é discutir a importância da integração de dados de diferentes fontes para melhor representar o universo da pesquisa, destacando como a restrição de fontes impacta nos resultados de análises bibliométricas.

Como objeto de estudo foi eleita a Zika, tendo como foco a produção brasileira na área, através da seleção de artigos de periódicos com pelo menos um autor brasileiro.

## 2 METODOLOGIA

Para facilitar o entendimento do trabalho, a metodologia se apresenta em duas etapas centrais, sendo a primeira referente a coleta de dados e a segunda ao tratamento de dados e integração das bases. As etapas são detalhadas abaixo.

### 2.1 COLETA DE DADOS

Para a coleta da produção científica foram utilizadas quatro bases de dados: duas fontes gerais, a *WoS* e a *Scopus*; uma fonte específica da área de estudo, o *PubMed* e uma fonte de pesquisa de informação nacional, o *SciELO*. A estratégia utilizada em todas as bases foi o uso dos termos (*zika or zikv or zikav*) no campo correspondente à título, resumo e palavras-chave, publicados no período de 2014 a 2018<sup>1</sup>. Nas quatro fontes informacionais, os registros foram refinados de forma a recuperar apenas artigos de periódicos, com pelo menos um dos autores com instituição de origem localizada no Brasil.

### 2.2 TRATAMENTO DE DADOS E INTEGRAÇÃO DAS BASES

Os dados coletados foram tratados com o software *VantagePoint* para eliminar duplicatas, fazer refinamentos, padronização de campos e integrar as fontes. Inicialmente, foram eliminadas eventuais duplicatas presentes

---

1 A busca foi realizada em janeiro de 2018.



nas próprias bases, o que pode ocorrer devido a inconsistências como, por exemplo, um mesmo artigo indexado como *pré-print* e publicado.

Em seguida, foram identificados campos comuns entre as bases para integração: título, autores, periódico, ano de publicação e países dos autores. Foram também utilizados campos para facilitar identificação de duplicatas entre as bases: DOI, volume, número e página inicial.

Em função das grandes discrepâncias no modo de indexação da informação de instituição do autor, optou-se por não utilizar esse campo no estudo. Foi observada ainda a necessidade de padronização de alguns campos para harmonização entre as bases, sendo estes: título do artigo, nome do periódico, país e autor. Como exemplo, pode-se citar o fato dos títulos dos artigos oriundos da base PubMed possuírem pontuação ao final do texto do título sendo necessário seu processamento, bem como a necessidade de padronizar nomes de países na base SciELO envolvendo casos como o Brasil para o qual pode ser encontrada a grafia em língua inglesa (Brazil). Os nomes de periódicos também se encontram com diferentes formas nas bases: no *PubMed* se encontra abreviado e na *WoS* por extenso, por exemplo. No que se refere aos nomes dos autores, algumas bases utilizam iniciais com pontos e outras sem pontos.

A partir da definição e padronização dos campos comuns, foi realizada então a integração das bases, duas a duas. Inicialmente foram integradas *WoS* e *Scielo*, seguidas de *PubMed* e *Scopus*. Esses dois conjuntos foram então integrados de modo a se obter um conjunto único composto dos registros das quatro bases de dados utilizadas.

Em seguida, foram identificadas e removidas as duplicatas no conjunto final para avaliação dos registros únicos por meio de sucessivas rodadas de comparação de campos, sobretudo DOI. Como esse dado não estava presente em todos os registros, foi necessária uma segunda rodada em que se utilizou como critério que fossem exatamente iguais: o título do artigo, o ano de publicação, o nome, número e volume do periódico. Uma última verificação foi feita ainda através da análise individual de títulos de artigos iguais para a eliminação de registros duplicados.

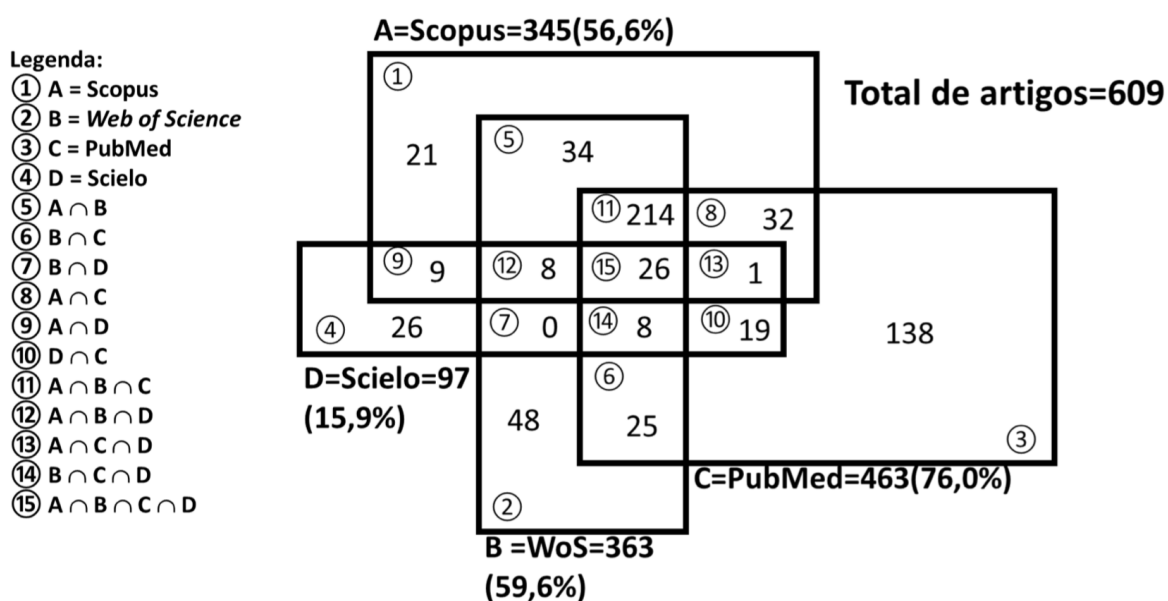


Por fim, foi realizado o tratamento dos dados referentes aos nomes dos autores, visando a comparação entre a posição dos principais autores considerando cada base individualmente e as quatro bases integradas. Para tanto, foi feito o uso de ferramentas de inteligência artificial do software *VantagePoint* para uniformizar o nome dos autores em cada base individualmente, com a conferência manual dos top 20 autores, verificando se haveria alguma grafia não identificada automaticamente. Nessa análise, foi desconsiderada a base SciELO, devido a sua baixa representatividade no conjunto total.

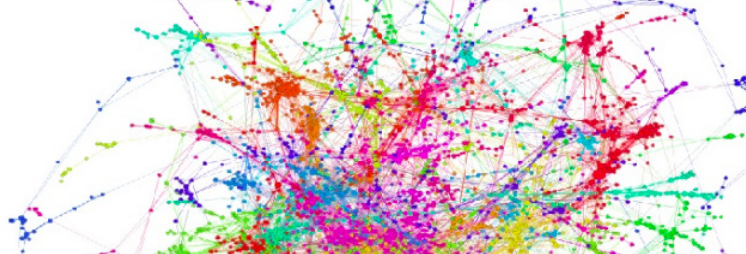
### 3 RESULTADOS E DISCUSSÃO

Foram encontrados um total de 609 artigos publicados no período do estudo, distribuídos pelas quatro bases de dados. Na *PubMed* foram localizados 463 artigos, na *WoS* 363, na *Scopus* 345 e no SciELO 97 documentos. No entanto, se observa alguma sobreposição entre esses documentos, conforme pode ser visto no Diagrama de Venn apresentado na Figura 1. Esse diagrama mostra as interseções entre as quatro bases de dados para facilitar a compreensão.

**FIGURA 1 - NÚMERO DE ARTIGOS POR BASE DE DADOS E SUAS INTERSEÇÕES**



**Fonte:** Elaborado pelos autores, 2018.



Em uma análise bibliométrica “padrão”, em que se usa somente a *WoS*, haveria a perda de cerca de 40% de artigos na análise. No caso de uso da *Scopus*, a perda seria de cerca de 43%. Mesmo se fosse utilizada uma base específica do tema de estudo, a *Pubmed*, ainda haveria a falta de cerca de 30% de documentos na temática. Se fossem usadas a *WoS* integrada com a *Scopus*, atingiria um percentual de cerca de 70% do conjunto total. No entanto, a união da *WoS* com a *Pubmed* traria um resultado mais satisfatório, atingindo 90% dos resultados.

A Figura 1 apresentada também nos permite perceber que somente 26 artigos estão presentes em todas as bases e há 214 documentos que parecem ser mais relevantes por estar presentes em três bases.

A integração de bases limita os campos que podem ser analisados, mas há uma tendência das bases uniformizarem seus dados. Além disso, a sofisticação das ferramentas de linguagem natural permite o uso de mecanismos que facilitam o tratamento de dados e possibilitam aproximações para análises bibliométricas confiáveis. Para fins de ilustração, os dados de nomes de autor foram padronizados para comparar a posição em que se encontram em cada base individualmente e no conjunto integrado das quatro bases. O Quadro 1 apresenta a lista dos autores com 10 ou mais publicações indexadas na *WoS* e sua posição nesse conjunto. As colunas seguintes apresentam o posicionamento desses autores por número de publicações nas bases de dados *Scopus* e *Pubmed*, bem como o posicionamento considerando as quatro bases integradas.



### QUADRO 1 – COMPARATIVO DA POSIÇÃO DOS AUTORES MAIS PRODUTIVOS DENTRE AS BASES DO ESTUDO

| Autor [número de artigos na WoS] | Posição na base |        |        |         |
|----------------------------------|-----------------|--------|--------|---------|
|                                  | WoS             | Scopus | Pubmed | 4 bases |
| Cordeiro, Marli Tenorio          | 1               | 1      | 1      | 1       |
| Bispo de Filippis, Ana Maria     | 2               | 3      | 2      | 4       |
| van der Linden, Vanessa          | 2               | 2      | 5      | 5       |
| Brasil, Patricia                 | 3               | 4      | 2      | 3       |
| Nogueira, Mauricio Lacerda       | 4               | 8      | 4      | 6       |
| Tanuri, Amilcar                  | 5               | 6      | 3      | 7       |
| Vasconcelos, Pedro F C           | 5               | 5      | 2      | 2       |
| de Oliveira, Wanderson Kleber    | 6               | 9      | 12     | 11      |
| Lourenco-de-Oliveira, Ricardo    | 7               | 8      | 7      | 9       |

**Fonte:** Elaborado pelos autores, 2018.

Observa-se que há uma variação importante no posicionamento dos autores nas diferentes bases de dados. Por exemplo, o autor que tem a segunda maior produção, considerando as quatro bases de dados, se encontra em 5º lugar tanto na *WoS* como na *Scopus*. Em estudos de indicadores para financiamento de pesquisas é usual se considerar apenas uma base de dados, o que pode trazer uma repercussão importante no investimento nas pesquisas de um autor com uma produtividade maior do que a representada em uma única base. Por outro lado, um autor que se posiciona em, por exemplo, 2º lugar na *WoS*, quando considerada as quatro bases integradas, está localizado na 5ª posição.

Esses exemplos indicam que é necessário estimular ações para atender as recomendações do Manifesto de Leiden (HICKS, 2015). Dentre as 10 recomendações sugeridas para orientar o uso de métricas em avaliação da pesquisa, a terceira refere-se à importância de proteger a excelência da pesquisa desenvolvida e publicada localmente. Em geral os estudos bibliométricos tomam como base as publicações em inglês, mas em muitas áreas, como a Saúde Pública, por exemplo, há inúmeras pesquisas que tem dimensão regional ou nacional. O pluralismo da sociedade tende a



ser suprimido para criação de artigos de alto impacto, em periódicos de língua inglesa. Métricas desenvolvidas a partir de literatura de alta qualidade em outros idiomas são úteis para identificar e premiar a excelência na pesquisa relevante localmente. O uso de diferentes bases de dados tende a minimizar essa questão, e a integração das bases é um caminho importante a ser investigado e novas metodologias desenvolvidas nesse sentido.

## REFERÊNCIAS

ARAÚJO, K. M.; SILVA, C. H.; GUIMARÃES, M. C. S. Produção científica e doenças emergentes: o caso da Zika. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. **Anais...** Salvador: ANCIB.

ARAÚJO, K. M.; SILVA, C. H.; GUIMARÃES, M. C. S.; LINS, R. A.; ASSEF NETO, R. C. de S. A produção científica sobre zika em periódicos de acesso aberto. **RECIIS – Rev. Eletron. Comun. Inf. Inov. Saúde**, v. 11, n. supl., out.-dez. 2017.

GHENO, E. M.; SCHÜLER-FACCINI, L.; SOUZA, D. O.; CALABRÓ, L. Panorama e características da produção científica brasileira e internacional sobre Zika vírus. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais...** São Paulo: USP, 2016. p. A26.

GHENO, E. M.; SCHÜLER-FACCINI, L.; SOUZA, D. O.; CALABRÓ, L. Vírus Zika: produção e colaboração científica brasileira. In: REUNIÃO ANUAL DA SBPC, 69., 2017, Belo Horizonte. **Anais...** Belo Horizonte: UFMG, 2017. p. 1-4.

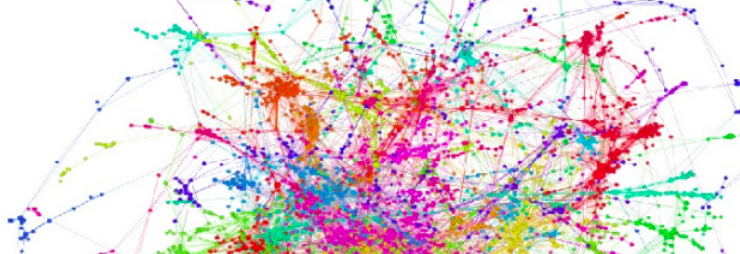
HICKS, D. et al. The Leiden Manifesto for research metrics. **Nature**, v. 520, n. 7548, p. 429, 2015.

NUNES, M. L.; CARLINI, C. R.; MARINOWIC, D.; KALIL NETO, F.; FIORI, H. H.; SCOTTA, M. C.; ZANELLA, P. L. Á.; SODER, R. B.; COSTA, J. C. Microcephaly and Zika virus: a clinical and epidemiological analysis of the current outbreak in Brazil. **Jornal de Pediatria**, v. 92, n. 3, p. 230-240, 2016 .

RIBEIRO, K. da C.; PAULO, A. C. L.; DANTIER, R. M. P.; BATISTA, F. B.; MIRANDA, G. M. Zika vírus: indicadores bibliométricos das publicações na base Scopus. In: CONGRESSO DE INTERDISCIPLINARIDADE DO NORDESTE

6° EBBC

Rio de Janeiro  
17 a 20 de julho



FLUMINENSE, 1., 2016, Itaperuna. **Anais...** Itaperuna: Instituto Federal Fluminense, 2016.