Correspondence on 'Mining social media data to investigate patient perceptions regarding DMARD pharmacotherapy for rheumatoid arthritis'

In their recent article, Sharma *et al*[1] posed an interesting research question where they used social media data to examine patients' perceptions of disease modifying anti-rheumatic drug (DMARD) pharmacotherapy for rheumatoid arthritis (RA). The authors used the web analytics platform Treato, which provides data from multiple social media sites. They conducted a sentiment analysis towards DMARDs and identified patient beliefs within the positive and negative sentiment.

While we agree with the authors' view that social media provides a valuable data source to explore patients' perspectives, the study design leaves uncertainties about the authors' results and conclusions. Our primary concern is about the claim that the study establishes causality based on research data of uncertain, potentially low quality and correlations that denote the association between two quantitative variables.

First, the type of data that was included in the data analysis is unclear. The authors did not describe a clear data search strategy (eg, keywords, phrases, hashtags), and they did not provide evidence about the relevance of the included data to the research question. However, the importance of a transparent and adequately focused search, filtering and quality evaluation of social media data has been pointed out by Kim *et al* and others.[2–4] Without it, the conclusions researchers draw may be biased or misleading. The complementary use of qualitative research methods designed to describe and interpret complex phenomena such as individuals' views and beliefs could have helped to generate some evidence in this regard.

Second, the authors relied on Treato's algorithm to identify patients' self-reported experiences. Still, they did not offer any evidence that suggests the algorithm's accuracy to identify patient-generated posts reliably. It is well-documented that human-like automated accounts by commercial and other interest groups pollute public social media data.[5–9] In this study, it remains unclear how many of the analysed posts originated from people with RA. Using qualitative research methods and tools such as Botometer[10] to identify patient's accounts and related posts would have strengthened the study's internal validity.

Third, the authors attempted to validate the results of Treato's sentiment analysis by manually assigning sentiments to 200 posts and comparing their codings with the algorithm's results. They used Cohen's Kappa coefficient for measuring the interobserver agreement. We wondered why the authors considered moderate values between k=0.49 for csDMARDs and k=0.52 for b/tsDMARDs sufficient, considering that the study topic has clinical relevance. Most texts recommend a minimum of 80% inter-rater agreement as acceptable.[11] The authors rightly point out that a sentiment analysis standard does not exist. However, we see this as an argument for more cautionary results' interpretation, especially when a study relies heavily on computer-assisted data collection and analysis methods.

Lastly, the authors intended to describe patients' beliefs associated with the positive and negative sentiments towards DMARDs, but it is unclear how these 'beliefs' were determined. There is no description of a priori codes that may have been used to identify complex concepts, such as 'efficacy'.[12 13] More importantly, the paper claims to have established causality. According to the authors, 'The two most common reasons for a positive post were DMARD efficacy and lack of side effects. Conversely, the most common reasons for a negative post were lack of efficacy and side effects'. However, as far as we can tell from the method description, the algorithm established correlations of keywords. Deriving causality directly from correlations is invalid across scientific disciplines.[14] For instance, a hypothetical statement such as 'I had frequent headaches with the previous therapy, and this made me passive and depressed, so I am trying now to switch to (drug of interest)' has all the relevant keywords for side effect, the drug of interest and negative sentiment. Yet, it does not express causality between the drug of interest and the side effect. Without further discussion or additional evidence, the authors' causality claims are not substantiated.

We acknowledge that the authors are in line with a current trend and focus on big data studies, which pose unique challenges. The focus on big data raises fundamental questions about the characteristics of a reliable dataset and the extraction of valid and meaningful insight.[15] We find that the paper would greatly benefit from reflections on these questions.

In closing, we hope that this correspondence will be received as constructive criticism to help move this research field forward. Social media-based patient-generated data have the potential to offer a new type of medical evidence. The ability to take full advantage of it will require technical advance, critical thinking and the re-evaluation of conventional evidence standards. In the meantime, we need to be aware of the limitations and challenges of using social media to capture patients' perspectives and interpret the data with the appropriate caution.

Katja Reuter ,[1] **Elena Rocca**[2]

[1]Department of Public Health & Preventive Medicine, SUNY Upstate Medical University, Syracuse, New York, USA
[2]Centre for Applied Philosophy of Science, Norwegian University of Life Sciences, Ås, Akershus, Norway

Correspondence to Dr Katja Reuter, Department of Public Health & Preventive Medicine, SUNY Upstate Medical University, Syracuse, NY 13210-2306, USA; reuterk@upstate.edu

Twitter Katja Reuter @dmsci

Check for updates

Linked

► https://doi.org/10.1136/annrheumdis-2020-219815

REFERENCES

1. Sharma C, Whittle S, Haghighi PD, *et al*. Mining social media data to investigate patient perceptions regarding DMARD pharmacotherapy for rheumatoid arthritis. *Ann Rheum Dis* 2020;79:1432–7.
2. Kim Y, Huang J, Emery S. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, Infodemiology and digital disease detection. *J Med Internet Res* 2016;18:e41.
3. Reuter K, Lee D. Evaluating patients' perspectives on social media: the importance of clearly reporting data search, cleaning and processing. *Br J Dermatol* 2019;181:222.
4. McDonald L, Malcolm B, Ramagopalan S, *et al*. Real-World data and the patient perspective: the promise of social media? *BMC Med* 2019;17:11.
5. Allem J-P, Ferrara E, Uppu SP, *et al*. E-Cigarette surveillance with social media data: social Bots, emerging topics, and trends. *JMIR Public Health Surveill* 2017;3:e98.
6. Ferrara E, Varol O, Davis C. The rise of social Bots, 2016. Available: https://cacm.acm.org/magazines/2016/7/204021-the-rise-of-social-bots/fulltext [Accessed 9 Jul 2019].
7. Shao C, Ciampaglia GL, Varol O, *et al*. The spread of low-credibility content by social bots. *Nat Commun* 2018;9:4787.
8. Mønsted B, Sapieżyński P, Ferrara E, *et al*. Evidence of complex contagion of information in social media: an experiment using Twitter bots. *PLoS One* 2017;12:e0184148.
9. Pozzar R, Hammer MJ, Underhill-Blazey M, *et al*. Threats of Bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. *J Med Internet Res* 2020;22:e23021.
10. Davis CA, Varol O, Ferrara E. BotOrNot: A System to Evaluate Social Bots. arXiv [cs.SI], 2016. Available: http://arxiv.org/abs/1602.00975
11. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22:276–82.
12. Sheard L, Marsh C, Mills T. *Using patient experience data to develop a patient experience toolkit to improve hospital care: a mixed-methods study*. NIHR Journals Library, 2019.
13. Ahmed F, Burt J, Roland M. Measuring patient experience: concepts and methods. *Patient* 2014;7:235–41.
14. Anjum RL, Mumford S. *Causation in science and the methods of scientific discovery*. Oxford: Oxford University Press, 2018.
15. Leonelli S. The challenges of big data biology. *Elife* 2019;8. doi:10.7554/eLife.47381. [Epub ahead of print: 05 Apr 2019].