Metaanalyses, Network Metaanalyses, and Systematic Reviews: The Perpetual Motion Machine All Over Again

The term *metaanalysis* was first used in the mid-1970s for describing methods designed to characterize and combine the findings of prior studies to increase statistical power, along with providing quantitative summary estimates, and to identify data gaps and biases[1]. (In this editorial I will use the term *metaanalysis* to encompass not only metaanalyses but also systematic reviews and network metaanalyses, because the issues I raise apply to all of them and their variations.) When applied to studies conducted with similar populations and methods, metaanalyses can be useful. However, this is not the case with many metaanalyses where the findings of studies that differ in important ways have been combined, prompting the comment that "they have mixed apples and oranges" — and sometimes "apples, lice, and killer whales — yielding meaningless conclusions"[1,2].

Combining the results of individual studies potentially increases the total number of participants, and this should mean increased statistical power, yet differences in participant demographics and study methods may actually lead to decreased power owing to variability in the patient characteristics[1]. This then leads to more difficulty in ascertaining the real effects.

Add to this the issue of unpublished research to potentially skew the conclusions, because positive findings get published more often than negative results, starting with the decision to submit them in the first place[3]. It has been reported that falsified data also make it into metaanalyses[4]. In one example authors showed that 46% of all metaanalysis publications had their conclusions changed by publications with falsified data and 32% of all the analyses had a considerable change in the outcome[5].

There has also been a surge in the number of metaanalyses published over the years. The rate of growth was significantly greater for metaanalysis at 4676% compared to randomized clinical trials (RCT) at 138% during the same time period[6]. Metaanalyses may help to synthesize and update the literature using valuable methods of evidence-based medicine;

however, only an estimated 3% of them are methodologically sound, nonredundant, and provide useful clinical information[7]. Although the optimal metaanalysis/RCT ratio has yet to be determined, an ever-increasing proportion of this literature may provide minimal value, which should precipitate a reappraisal of the foundations, production, and reporting of metaanalyses[6,8].

Many potential reasons for this trend of an exploding number of metaanalyses have been proposed, ranging from an actual need for updating accumulated evidence and hence the need for summarized data, to padding of resumes and journal citation statistics[9,10]. Others have also suggested that metaanalysis may serve as "easily publishable units or marketing tools"[11,12]. Even what is considered the gold standard for metaanalyses, Cochrane Reviews, has been shown to not meet its own standards in its reports, and the scandal around the firing of one of its founders should give pause to anyone who cares about the sanctity of scientific rigor[13,14]. These recent trends have led to questions about the purpose, quality, and credibility of most reviews as well as calls to abandon metaanalyses altogether, and that part of the responsibility falls on the journal editors and reviewers to make sure only good quality work gets published[15].

A potential problem for rheumatology and specifically for rheumatoid arthritis (RA) studies is how the changed classification criteria for RA will affect future recommendations when a metaanalysis is done looking at treatment options for RA. The main issue is that the new criteria published in 2010[16] have been shown, by us and others, to have decreased specificity, which of course leads to patients who do not have RA and have other diagnoses explaining their condition to be classified as having RA[17,18,19]. There have also been data suggesting that patients with RA classified using the 2010 criteria have less severe disease, respond better to treatments, and have improved remission rates[20]. Some have suggested that RA itself is changing. A far more plausible explanation is that the new criteria, by the way they were developed,

select for patients with milder disease and even patients who do not have RA to be enrolled into clinical trials, hence skewing the results. Imagine the confusion and likely incorrect conclusions that would be reached after a meta-analysis with these skewed results.

We seem to believe that more data from many patients, regardless of how the data were collected, analyzed, and reported, will answer many questions that a single, well-done trial would not. I disagree. A well-done study would probably tell us a lot more than 20 studies combined if a lot of them have methodological flaws, and very likely somewhat different types of patients studied, as mentioned before. It is much more straightforward to dissect a single study to really understand what the question asked was, how it was studied, and what the conclusions were than to try to interpret a metaanalysis where you do not know how the many potential issues listed above have affected the conclusions.

The reason for doing an RCT is not that it can someday be part of a metaanalysis. Maybe we should be more focused on the misplaced desire to keep pooling trials that probably should not be pooled to draw conclusions that should not be drawn. Each trial's only goal, in the case of drugs being tested, is to show if something works, yes or no, plus or minus, 1 or 0. All the other derivative conclusions are nice to have and can lead to further hypothesis development for the next study. RCT, however, are very good tools for saying that a certain medication works, and you should potentially offer it to a specific patient to see how that patient would do. Nothing more, nothing less.

I think the time has come to limit RCT and their conclusions to what was measured in that trial. The attempt to draw more than what these individual RCT can provide is the problem. We are constantly looking for the shortcuts that are not there. This is no different from all the attempts at personalized medicine for complex conditions, such as hypertension, diabetes, or RA, where it has been very robustly demonstrated that predicting outcomes at a single-patient level will very likely never be achieved[21]. As Roberts, *et al* stated, "Thus, our results suggest that genetic testing, at its best, will not be the dominant determinant of patient care and will not be a substitute for preventative medicine strategies incorporating routine checkups and risk management based on the history, physical status and life style of the patient … Recognition of these merits and limits … can minimize unrealistic expectations and foster fruitful investigations[21]." This love of trying to draw simple conclusions that would be applicable to all patients seems similar to 17th-century attempts to develop a perpetual motion machine. Everybody really loved the idea and wanted it to be possible (similar to the enthusiasm for individualized genetic testing or personalized medicine attempts), but you cannot break the first law of thermodynamics. Hence, there will never be a perpetual motion machine.

The latest incarnation of this kind of wishful thinking is related to artificial intelligence and the "era of big data," which can be thought of as the next step in the metaanalyses movement[22]. I remember the days when all would be solved if we only could sequence the whole human genome. We did, and learned a lot about diseases, but we found no insights into predicting diseases, best treatments, or outcomes in an individual patient with a common disease, which is what most people have and what most doctors try to treat. I would respectfully suggest that while we still can, we should try going back to what I will call "small" data, where only a few, well-done studies, with the aim of answering a hypothesis-driven question, are taken seriously and used in making treatment recommendations and decisions, because I do not know of more serious work for a doctor than taking care of an individual patient.

Benjamin Franklin said when he was a young man, "Lose no time; be always employ'd in something useful; cut off all unnecessary actions." Maybe it is time we applied this to our approach to most metaanalyses.

YUSUF YAZICI, MD,
Clinical Associate Professor of Medicine,
New York University, School of Medicine,
New York, New York, USA.
Address correspondence to Dr. Y. Yazici, 333 East 38th St., New York, New York 10016, USA. E-mail: yusuf.yazici@nyumc.org

REFERENCES

1. Barnard ND, Willett WC, Ding EL. The misuse of meta-analysis in nutrition research. JAMA 2017;318:1435-6.
2. Eysenck H. Meta-analysis squared: does it make sense? Am Psychol 1995;50:110-1.
3. Ioannidis JP. Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? Philos Ethics Humanit Med 2008;3:14.
4. Garmendia CA, Bhansali N, Madhivanan P. Research misconduct in FDA-regulated clinical trials: a cross-sectional analysis of warning letters and disqualification proceedings. Ther Innov Regul Sci 2018;52:592-605.
5. Garmendia CA, Nassar Gorra L, Rodriguez AL, Trepka MJ, Veledar E, Madhivanan P. Evaluation of the inclusion of studies identified by the FDA as having falsified data in the results of meta-analyses: the example of the apixaban trials. JAMA Intern Med 2019;179:582-4.
6. Niforatos JD, Weaver M, Johansen ME. Assessment of publication trends of systematic reviews and randomized clinical trials, 1995 to 2017. JAMA Intern Med 2019 Jul 29 (E-pub ahead of print).
7. Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 2016;94:485-514.
8. Møller MH, Ioannidis JPA, Darmon M. Are systematic reviews and meta-analyses still useful research? We are not sure. Intensive Care Med 2018;44:518-20.
9. Patsopoulos NA, Analatos AA, Ioannidis JP. Relative citation impact of various study designs in the health sciences. JAMA 2005;293:2362-6.
10. Qadir XV, Clyne M, Lam TK, Khoury MJ, Schully SD. Trends in published meta-analyses in cancer research, 2008-2013. Cancer Causes Control 2017;28:5-12.
11. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven

systematic reviews a day: how will we ever keep up? PLoS Med 2010;7:e1000326.

12. Siontis KC, Ioannidis JPA. Replication, duplication, and waste in a quarter million systematic reviews and meta-analyses. Circ Cardiovasc Qual Outcomes 2018;11:e005212.

13. Franco JVA, Garrote VL, Escobar Liquitay CM, Vietto V. Identification of problems in search strategies in Cochrane Reviews. Res Syn Methods 2018;9:408-16.

14. Gøtzsche PC. Cochrane — no longer a collaboration. [Internet. Accessed August 15, 2019.] Available from: blogs.bmj.com/bmj/2018/11/08/peter-c-gotzsche-cochrane-no-longer-a-collaboration

15. Wallach JD. Meta-analysis metastasis. JAMA Intern Med 2019 Jul 29 (E-pub ahead of print).

16. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO 3rd, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League against Rheumatism collaborative initiative. Ann Rheum Dis 2010;69:1580–8.

17. Kennish L, Labitigan M, Budoff S, Filopoulos MT, McCracken WA, Swearingen CJ, et al. Utility of the new rheumatoid arthritis 2010 ACR/EULAR classification criteria in routine clinical care. BMJ Open 2012;2:e001117.

18. Kaneko Y, Kuwana M, Kameda H, Takeuchi T. Sensitivity and specificity of 2010 rheumatoid arthritis classification criteria. Rheumatology 2011;50:1268–74.

19. van der Linden MP, Knevel R, Huizinga TW, van der Helm-van Mil AH. Classification of rheumatoid arthritis: comparison of the 1987 American College of Rheumatology criteria and the 2010 American College of Rheumatology/European League Against Rheumatism criteria. Arthritis Rheum 2011;63:37–42.

20. Burgers LE, van Nies JA, Ho LY, de Rooy DP, Huizinga TW, van der Helm-van Mil AH. Long-term outcome of rheumatoid arthritis defined according to the 2010-classification criteria. Ann Rheum Dis 2014;73:428-32.

21. Roberts NJ, Vogelstein JT, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE. The predictive capacity of personal genome sequencing. Sci Transl Med 2012;4:133ra58.

22. Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype? JAMA 2019;321:2281-2.