OA Evidence-Based Medicine
Open Access

Systematic Review

Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation?

JM Thomas[1*]

Abstract

Introduction

Systematic reviews, the foundation of much Evidence-Based medicine, are suffering from increasing 'data deluge': reviewers often need to manually assess many thousands of titles and abstracts to determine their relevance. Automation has been advanced as a potential solution; but given that its efficacy was first demonstrated in 2006, why is it not yet widely used? The Diffusion of Innovations framework by EM Rogers is used to structure an exploration of why this might be the case.

Discussion

According to Rogers, five characteristics affect the rate of adoption of innovations: those perceived as having greater relative advantage, compatibility, trialability and observability, and less complexity, will be adopted more rapidly than others. The relative advantage of automation has been demonstrated empirically, though usually in narrowly focused reviews in clinical areas, rather than more challenging areas for automation, such as public health. Detailed methods and procedures for their use have yet to be established, addressing transparency, replicability and reporting practices. Although issues concerning the compatibility of new technology with existing infrastructure are probably surmountable, the use of automation may challenge contemporary notions of what constitutes a systematic search and how publication bias is addressed using sensitive search techniques. The remaining factors are interrelated. The technologies are complex, both to understand and to deploy. This affects the trialability of automation: technical expertise is required and there are thus few opportunities for reviewers to observe others using these technologies.

Conclusion

Further technical and empirical work is needed where systematic reviewers work with information and computer scientists to develop solutions which have a demonstrative relative advantage and which are clearly compatible with the needs of systematic reviewers and their users. Such work may have a significant role to play in addressing the deluge of new research publications which threaten to overwhelm systematic review processes.

Introduction

Systematic reviews, the foundation of much Evidence-Based medicine, are suffering from increasing 'data deluge'. With 'seventy-five trials and 11 systematic reviews' being published every day, a figure which is now almost certainly an underestimation, the burden of identifying relevant studies is escalating[1].

In order to limit the number of studies to screen, reviews (particularly in health technology assessments and other rapid reviews) tend to adopt pragmatic and relatively specific strategies to searching–even though relevant research is probably missed because of this[2]. This problem is being compounded by an increase in the number of databases to search, especially as recent work has suggested that there is an inbuilt North-American bias in many major bibliographic databases (e.g. PubMed), and that a wide range of smaller databases need to be searched in order to identify research for reviews that aim to maximise external validity (e.g. for use in a UK context[3]).

Unfortunately, the specificity of searches conducted in bibliographic databases is low. Reviewers often need to manually assess many thousands of irrelevant titles and abstracts in order to identify the much smaller number of relevant ones[4], a process known as screening.

Reviews that address complex health issues, or that deal with a range of interventions (e.g. a typical public health review might be concerned with 'interventions to promote physical activity') are often those that have the most challenging numbers of items to screen; and reviews for which the traditional 'PICOS' (Participants, Intervention(s), Comparator(s), Outcome(s) and Study design(s)) search terms do not distinguish between eligible and illegible studies are likewise likely to be burdened with significant screening workload[5].

Given that an experienced reviewer can take between 30 s and several minutes to evaluate a citation[6], the work involved in screening 10,000 citations is considerable, and the screening burden in some reviews is considerably higher than this.

Increasing the use of automation at this phase of the review process has been advanced as a potential solution to the problem of data deluge[6–9]. Machine learning techniques, it is claimed, are able to reduce the screening burden by more than 50%, without resulting in any studies being 'lost' to the review. Given that the first demonstration of machine learning efficacy in this area was published in 2006, it is perhaps

*Corresponding author
Email: j.thomas@ioe.ac.uk

[1] EPPI-Centre, Social Science Research Unit, Institute of Education, London

Competing interests: None declared. Conflict of interests: None declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

FOR CITATION PURPOSES: *Thomas JM. Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation? OA Evidence-Based Medicine 2013 Oct 21;1(2):12.*

surprising that few reviews report making use of these new technologies. Is this due to there being barriers to the take-up of these techniques, or are the technologies themselves not yet ready?

The classic Diffusion of Innovations framework by EM Rogers is used to structure the following exploration of these issues, facilitating both an overview of current automation approaches and an assessment of their readiness for use in 'live' reviews.

Discussion

The author has referenced some of its own studies in this review. The protocols of these studies have been approved by the relevant ethics committees related to the institution in which they were performed.

Challenges in the diffusion of this innovation

First published in 1962, and in its 5th edition in 2003, Everett Rogers' Diffusion of Innovations articulates a theory which explains how and why new ideas and technologies spread through communities[10].

Of particular interest in this paper are the five attributes of innovations which help us to understand the current state of the field, the reasons why this innovation does not appear to be diffusing and what we might do in the future to advance knowledge and practice.

According to Rogers, innovations have five characteristics which affect their rate of adoption; those which are perceived as having greater relative advantage, compatibility, trialability and observability, and less complexity, will be adopted more rapidly than others [Kindle location (KL):610][10].

Paying particular attention to relative advantage and compatibility, I now examine each of these characteristics in relation to the adoption of (semi-) automation for reducing screening workload in systematic reviews.

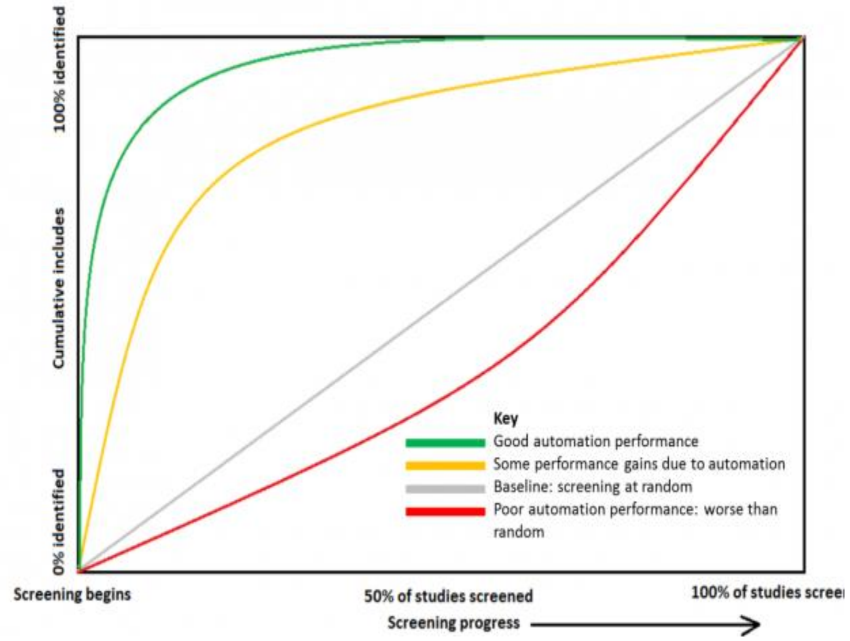

Figure 1: Changing the screening process.

Relative advantage

This is the degree to which an innovation is perceived as better than the idea it supersedes [KL: 586]. There is little doubt that some impressive results have been reported which describe how text mining can accelerate screening; but how are these results perceived and what is the comparator?

Screening is usually presented as a linear process in which a number of citations need to be manually checked for relevance to the review question. Figure 1 summarises this, with the process of screening each study manually being presented along the x-axis (0–100%) and the cumulative number of relevant studies ('includes') identified along the y-axis (0–100%). In a traditional review, where citations are screened essentially at random, we expect to see relevant studies identified in proportion to the number screened: depicted by the grey diagonal line.

Automation does not replace manual work, but aims to reduce it (hence, it should properly be referred to here as 'semi-automation'), by focusing manual effort on the most relevant citations, aiming to identify 100% of eligible studies as quickly as possible. The green line shows one possible result, with relevant studies identified at a much quicker rate and 100% of them found by the time about half of the citations have been examined. In theory, the reviewer can then discard the remaining 50% of citations, safe in the knowledge that they are not relevant to their review. Such results have been reported by several teams.

Wallace et al.[6] report that their technique might have reduced screening effort by between 40% and 50% in three reviews[6]; and by between 67% and 92% in four examples of review updates[11].

Cohen and colleagues[9,12] present similarly promising results, and at least three groups are building systems which use text mining to facilitate the retrieval of studies in reviews Stanley et al.[13], Thomas et al.[14] and Yang et al.[15].

These results are achieved through a process known as 'active learning'[6], which is illustrated in Figure 2. Briefly put, 'active learning' is an iterative process whereby the accuracy of the predictions made by the machine are improved through interaction with

Competing interests: None declared. Conflict of interests: None declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

users (reviewers). When used in a review, active learning involves the reviewer screening a small number of studies manually; the machine then 'learns' from these decisions and generates a list of citations for the reviewer to look at next. This cycle continues, with the number of reviewer decisions growing, until a given stopping criterion is reached and the process ends (e.g. the reviewer has identified all the relevant studies they had expected; they have run out of time and they have screened all the studies manually). The mechanism for generating the list of studies to be examined manually is under active consideration[16].

On the face of it then, it is possible to get good results from semi-automating the screening process, and some people are sufficiently convinced to invest in systems to support it. However, the number of evaluations is small and they rely almost exclusively on retrospective 'simulation' studies, where data from completed reviews are re-analysed and optimum conditions for the text mining tools are identified.

Questions remain in terms of being able to predict how well the tools may perform in other situations; they may, for example, obtain a curve similar to the yellow line in Figure 1. This is similar to the green line to begin with, but importantly, does not identify every relevant study until towards the end of the screening process. In this example, while more relevant studies are identified earlier in the process than usually is the case, it is still necessary to screen all citations manually in order to find them all.

At the moment, it is difficult to evaluate the relative advantage of simply identifying relevant studies earlier in the process (there are no evaluations), but one might assume there would be some, since the process of full-text retrieval, screening and data extraction could begin earlier in the review.

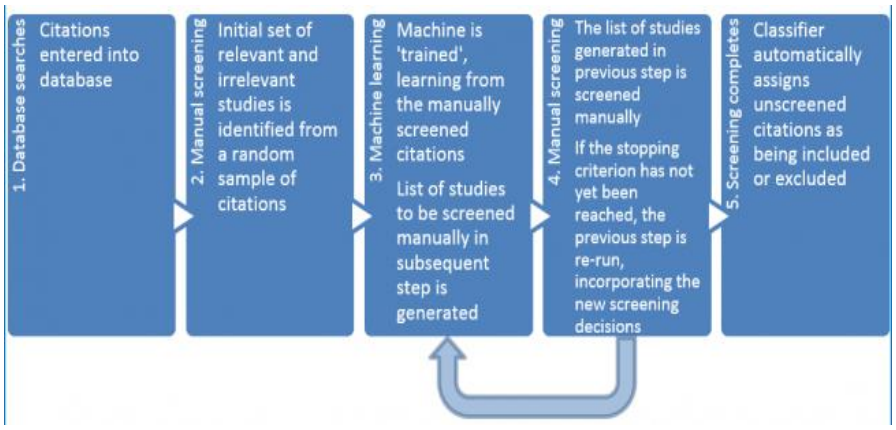
1. Database searches
Citations entered into database
2. Manual screening
Initial set of relevant and irrelevant studies is identified from a random sample of citations
3. Machine learning
Machine is 'trained', learning from the manually screened citations
List of studies to be screened manually in subsequent step is generated
4. Manual screening
The list of studies generated in previous step is screened manually
If the stopping criterion has not yet been reached, the previous step is re-run, incorporating the new screening decisions
5. Screening completes
Classifier automatically assigns unscreened citations as being included or excluded

Figure 2: Active learning process.

As mentioned above, current practice when constructing a search is to tailor the number of studies downloaded from databases to fit the time and resource available to look at them. The Cochrane Handbook illustrates this balance in saying that 'time and budget restraints require the review author to balance the thoroughness of the search with efficiency in use of time and funds…'[17]. The need to change practice and adopt a new way of screening using automation may therefore not be apparent, as the number of studies potentially eligible to download increases, current practice adapts to this by making searches more precise; thus, the relative advantage of doing larger searches, but using automation to manage the ensuing screening burden, is difficult to assess.

In addition to the above, questions of transparency, replicability and bias deserve a mention. Since automation methods depend upon a sample of relevant studies from which to 'learn', if the sample is biased in some way, there is the possibility that the process will systematically fail to find certain studies. In addition, few papers have explored the possibility of replicating the process of another group's semi-automated screening; and, nor are there agreed standards for reporting the results of screening in this way.

Compatibility

This is the degree to which an innovation is perceived as being consistent with the existing values, past experiences, and needs of potential adopters [KL: 586]. There are two dimensions to explore here in relation to compatibility: conceptual and infrastructural.

In terms of infrastructure, organisations which undertake many systematic reviews have established procedures and information systems. They either use 3rd party systematic review software, such as Covidence, Distiller SR, Dr Evidence or EPPI-Reviewer etc.; a reference management tool, such as Endnote; or standard 'Office' software such as Excel or Access. Reference management and generic office software will not support the type of automation discussed here (ever), and specialist systematic review software does not, on the whole, support automation; it is therefore clearly incompatible with current infrastructure.

The existence of specialist systematic review software, however, means that bespoke systematic review infrastructure could become compatible with this innovation, and this should therefore not be considered an insurmountable barrier.

Conceptually, though, automation may be incompatible with current understandings about what a systematic and unbiased search 'looks like'. Current practice often involves tailoring the sensitivity of a search to fit the resource available to screen the resulting citations. Thus, once all references have been retrieved, they become the 'universe' of eligible studies, and what is then needed is to check through them manually (Figure 3). Contrast this with what may be possible using automation: a reviewer might not restrict their search at all because they intend to reduce the screening burden using automation.

They then use the techniques mentioned above and only look at, say, 50% of the citations actually retrieved. In this example, the reviewer may have conducted a more sensitive search, and may actually have found more relevant studies; but they may also be able to estimate (using performance metrics) that they have found only 95% of the eligible studies that they had downloaded.

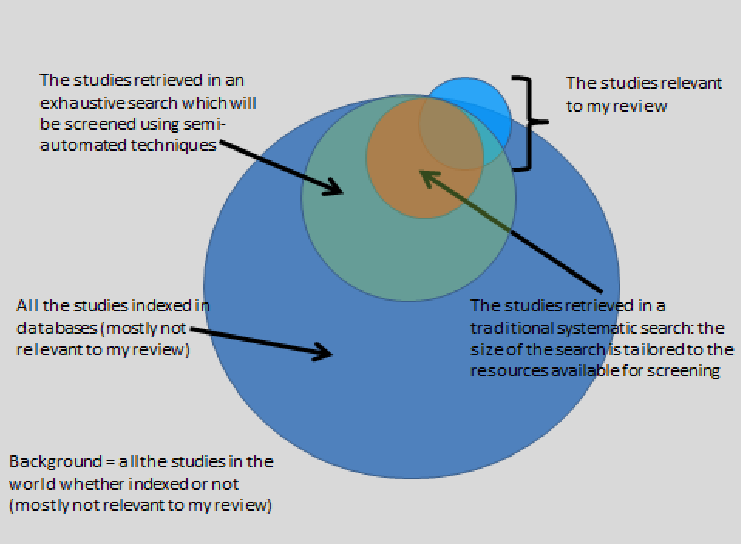
The studies retrieved in an exhaustive search which will be screened using semi-automated techniques
The studies relevant to my review
All the studies indexed in databases (mostly not relevant to my review)
The studies retrieved in a traditional systematic search: the size of the search is tailored to the resources available for screening
Background = all the studies in the world whether indexed or not (mostly not relevant to my review)

Figure 3: Conceptualising two approaches to searching (not to scale).

This would appear to be at odds with current notions of systematic searching, where the aim is to find every relevant study, and the degree to which tailoring the sensitivity of searches affects this is difficult to ascertain and rarely acknowledged.

In summary, there are open questions concerning both the infrastructural and conceptual compatibilities of automation in systematic reviews in the systematic review community, with methodological concerns probably being the greater obstacle to adoption.

Complexity

This is the degree to which an innovation is perceived as difficult to understand and use. Some innovations are readily comprehended by most members of a social system; others are more complicated and are adopted more slowly [KL:597]. Text mining / machine learning and the automation techniques that build on these technologies are at the forefront of current computer science and are the subject of interest and active development. It is a complex field with many competing approaches, its own language and accepted modes of working. It develops experimental software to demonstrate advances in thinking, but much software in this field is not made available for non-specialists to use in an applied way.

Moreover, both the ways that these technologies operate, and the highly technical language used to describe them means that non-specialists will find it difficult to understand how they work and how to fine-tune them for particular situations. Thus, without collaborative partnerships and/or some expertise in software engineering, most systematic reviewers will be unable to understand how semi-automation operates and how it might be used. In the longer run this may not be a problem, since the technologies might be amenable to being treated like a 'black box' with pre-set parameters; in the present though, their complexity is a barrier to understanding, and hence, acceptance.

Trialability

This is the degree to which an innovation may be experimented with on a limited basis [KL:597].Two challenges relate to the trialability of automation in systematic reviews: access to available and functional software, and an evidence-base to support its use in particular situations. The ability of a mainstream systematic reviewer to try out these tools in their own work is very limited. They require a high degree of technical expertise simply to select and run; and do not integrate into most systems and processes.

The barriers identified above in terms of complexity mean that most systematic reviewers will be unable to trial automation without assistance in terms of infrastructure and technical support.

Observability

This is the degree to which the results of an innovation are visible to others [KL:597]. In the light of the above challenges, it will be no surprise to find that observability is currently very low and is mostly restricted to journal articles (typically of a technical nature) and conference presentations. The problems discussed above in articulating a demonstrable relative advantage and in making the case for compatibility in conceptual terms, mean that it is particularly difficult for individuals to observe the results of this innovation and to assess for themselves its worth and viability.

Increasing diffusion and the rate of adoption

The above examination of the challenges of adopting these new technologies to reduce workload in screening suggests that their widespread use is some way off; and yet with the number of publications increasing ever faster, we need to find

For citation purposes: *Thomas JM. Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation? OA Evidence-Based Medicine 2013 Oct 21;1(2):12.*

ways of identifying studies for inclusion in systematic reviews in more efficient–but still reliable–ways. In order to better demonstrate the relative advantage of the technologies, we need far more evaluations of the use of semi-automation in a diverse range of systematic reviews, including both the (relatively straightforward) clinical and technical literature, as well as the (more challenging) social science and theoretical literature.

We also need evaluations of different approaches to searching and screening, which examine critically the current practice of limiting search sensitivity in accordance with the resources available and compare it with highly sensitive search strategies which utilise semi-automation.

Conceptual compatibility must be addressed through empirical work, to examine issues of potential bias, transparency and replicability; and a shift needs to take place where, conceptually, the locus of most search activity becomes the database of retrieved studies, rather than the many bibliographic databases where search options are often limited.

In addition, more consideration needs to be given to the purposes of systematic searching and the situations in which an exhaustive search is needed, or where a 'purposive' search might be more fit for purpose[5,18].

Finally, software tools which integrate into existing processes and systems require development. Besides overcoming infrastructural compatibility, such tools will address the problems identified in complexity and trialability.

Conclusion

Earlier in this paper I asked; is the lack of uptake of such technology due to there being barriers, or are the technologies themselves not yet ready? The answer seems to be both, although the technologies are progressing rapidly. An examination of the current state of the field in the light of the Diffusion of innovations theory reveals that we are as yet some distance from meeting the conditions where widespread adoption might occur.

Further technical and empirical work is needed where systematic reviewers work with information and computer scientists to advance the field and develop solutions which have a demonstrative relative advantage and which are clearly compatible with the needs of systematic reviewers and their users. Most importantly, this work needs to focus on the aforementioned areas of conceptual compatibility and relative advantage. Here, the barriers to the adoption of new technologies concern current conceptualisations of what constitutes a systematic search.

Due to the number of studies retrieved by sensitive searches, practice is often increasingly at variance with theory; yet, the danger is that opportunities to expedite certain processes may currently be missed because we have not yet advanced our understanding of how automation might be utilised in an unbiased, transparent and auditable way.

Once we have addressed these core conceptual challenges, then the other issues identified above are likely to be overcome quickly by increased community interest and enthusiasm.

Acknowledgment

This work was funded by a Medical Research Council (grant number MR/J005037/1). I would like to thank Alison O'Mara-Eves for her helpful comments and my other collaborators on the MRC grant, Sophia Ananiadou, John McNaught and Makoto Miwa.

Thanks are also due to participants at the #CochraneTech day in Quebec (September 2013) for discussion which prompted the main question posed in this paper.

References

1. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010 Sep;7(9):e1000326.
2. Watt A, Cameron A, Sturm L, Lathlean T, Babidge W, Blamey S, et al. Rapid reviews versus full systematic reviews: an inventory of current methods and practice in health technology assessment. Int J Technol Assess. Health Care. 2008;24(2):133–9.
3. Stansfield C, Kavanagh J, Rees R, Gomersall A, Thomas J. The selection of search sources influences the findings of a systematic review of people's views: a case study in public health. BMC Med Res Methodol. 2012 Apr;12(55).
4. Lefebvre C, Manheimer E, Glanville JPT. Searching for studies (chapter 6). In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.1.0 (updated March 2011). The Cochrane Collaboration; 2011.
5. Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synth Methods. 2013Dec;n/a:n/a.
6. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinform. 2010 Jan;11:55.
7. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Res Synth Methods. 2011 Mar;2(1):1–14.
8. Tsafnat G, Dunn A, Glasziou P, Coiera E. The automation of systematic reviews. BMJ. 2013 Jan;346:f139.
9. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc. 2006 Mar–Apr;13(2):206–19.
10. Rogers E. Diffusion of innovations. 5th ed. New York, NY: Free Press; 2003.
11. Wallace BC, Small K, Brodley CE, Lau J, Schmid CH, Bertram L, et al. Toward modernizing the systematic

Competing interests: None declared. Conflict of interests: None declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.

review pipeline in genetics: efficient updating via data mining. Genet Med. 2012 Jul;14(7):663–9.

12. Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. J Am Med Inform Assoc. 2009 Sep–Oct;16(5):690–704.

13. Ip S, D'Ambrosio C, Patel K, Obadan N, Kitsios GD, Chung M, Balk EM. Auto-titrating versus fixed continuous positive airway pressure for the treatment of obstructive sleep apnea: a systematic review with meta-analyses. Syst Rev. 2012 Mar;1:20.

14. Thomas J, Brunton J, Graziosi S. EPPI-Reviewer 4.0: software for research synthesis. London: EPPI-Centre Software, Social Science Research Unit, Institute of Education; 2010.

15. Yang JJ, Cohen AM, McDonagh MS., editors. SYRIAC: The systematic review information automated collection system a data warehouse for facilitating automated biomedical text classification. AMIA Annu Symp Proc. 2008 Nov:825–9.

16. Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. J Biomed Inform. 2013 Nov.

17. Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.1.0 (updated March 2011). The Cochrane Collaboration; 2011.

18. Brunton G, Stansfield C, Thomas J. Finding relevant studies. In: Gough D, Oliver S, Thomas J, editors. An introduction to systematic reviews. London: Sage; 2012; 107-134.

Competing interests: None declared. Conflict of interests: None declared.
All authors contributed to conception and design, manuscript preparation, read and approved the final manuscript.
All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.