Evaluation of language analysis to summarize the literature: a comparison to traditional meta-analysis in primary hip and knee surgery

Joshua Myszewski [1], Emily Klossowski,[2] Kristopher M Schroeder[3]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/rapm-2021-103261).

[1]School of Medicine and Public Health, University of Wisconsin-Madison, Madison, Wisconsin, USA
[2]College of Letters and Science, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA
[3]Anesthesiology, University of Wisconsin-Madison, Madison, Wisconsin, USA

Correspondence to
Joshua Myszewski, University of Wisconsin-Madison School of Medicine and Public Health, Madison, Wisconsin, USA; jmyszewski@wisc.edu

Received 15 October 2021
Accepted 22 November 2021
Published Online First 14 December 2021

ABSTRACT

Introduction Sentiment analysis, by evaluating written wording and its context, is a growing tool used in computer science that can determine the level of support expressed in a body of text using artificial intelligence methodologies. The application of sentiment analysis to biomedical literature is a growing field and offers the potential to rapidly and economically explore large amounts of published research and characterize treatment efficacy.

Methods We compared the results of sentiment analysis of 115 article abstracts analyzed in a recently published meta-analysis of peripheral nerve block usage in primary hip and knee arthroplasty to the conclusions drawn by the authors of the original meta-analysis.

Results A moderately positive outlook supporting the utilization of regional anesthesia for hip and knee arthroplasty was found in the 115 articles that were included for analysis, with 46% expressing positive sentiment, 35% expressing neutral sentiment, and 19% of abstracts expressing negative sentiment. This was well aligned with the conclusions reached by a previous meta-analysis of the same articles.

Discussion Sentiment analysis applied to the medical literature can rapidly evaluate large collections of published data and generate an impression of overall findings that are aligned with the findings of a traditional meta-analysis.

INTRODUCTION

As the volume of medical literature continues to grow, meta-analysis studies represent an increasingly used and important mechanism to distil the available published data and make conclusions based on larger patient sample sizes. The clinical utility of these analyses results is derived from strengthened supportive conclusions in those settings where study findings are congruent and in dismissal of findings that fail to be replicated across published studies. While meta-analyses represent a powerful mechanism to evaluate published study data, their conduct can be limited by the time required to locate and evaluate appropriate published studies, export data that are to be included, and perform a statistical analysis. As the number of questions and scenarios explored by clinicians expands, the time constraints associated with traditional meta-analysis studies may render these types of analyses unwieldy and impractical for practicing clinicians.

One potential option to address the growing size of literature on any topic is the utilization of artificial intelligence as an alternative or complement to traditional systematic review and meta-analysis. More specifically, a method such as sentiment analysis or other natural language processing methods could be used to quickly assess the findings of large groups of published studies and serve as a mechanism to generate further hypothesis.

Sentiment analysis is a type of artificial intelligence that can classify a body of text based on the qualitative sentiment (ie, the tone expressed) expressed within it and output either a categorical sentiment score (ie, positive, negative, neutral) or a numerical score which is on a spectrum where −1 is very negative and 1 is very positive.[1] This type of analysis has been applied in a variety of fields in the past for assessing large-scale trends such as those in social media[2] or bodies of clinical trial literature.[3] Furthermore, specific algorithms have been developed that facilitate sentiment analysis of clinical trial abstracts.[4] This presents an intriguing avenue for quick assessment of the qualitative statements made by the authors of a study. However, sentiment analysis as an adjunct or complement to other methods of systemic analysis of biomedical literature has not yet been explored.

With this in mind, we performed a sentiment analysis of the clinical trial abstracts used in a recent meta-analysis evaluating peripheral nerve block anesthesia use in primary hip and knee arthroplasty. We then compared the results of this sentiment analysis to what was reported in the meta-analysis. The goal of this study was to compare the results of sentiment analysis with the more exhaustive process of systematic review and explore its application as an adjunct to such methods of literature review.

METHODS

In this study, the sentiment (ie, the tone or level of support) of articles cited in a recent systematic review of the literature evaluating the use of peripheral nerve block analgesia for primary hip and knee arthroplasty was determined using GAN-BioBERT sentiment analysis.[4 5] These findings were then compared with those expressed in a recent meta-analysis evaluating analgesic techniques for total hip and knee arthroplasty.[5] This algorithm is publicly available under open-source license.[6]

Data collection

The abstracts of published manuscripts included in the systematic review by Memtsoudis *et al* that had abstracts available in the PubMed database were

Linked

► http://dx.doi.org/10.1136/rapm-2021-103381

Check for updates

To cite: Myszewski J, Klossowski E, Schroeder KM. *Reg Anesth Pain Med* 2022;**47**:151–154.

Table 1 Samples of phrases with each sentiment category and value as determined by GAN-BioBERT

Sample phrases	Sentiment classification
The results of this study were promising.	Positive
The results of this study showed that treatment was contraindicated.	Negative
There were no significant differences found between the treatment and control group.	Neutral

collected using the NCBI's Entrez E-Utilities API by identifying the associated PMID for each paper.[5] Manuscripts that either did not have abstracts available or were not indexed in PubMed were excluded.

Subgroup analyses

From the abstracts, several subgroup analyses were also performed. The subgroups included division by anesthesia technique, study type, surgery type (ie, knee vs hip arthroplasty), risk of bias, and number of patients in the study.

For the bias-based subgroups, studies with a bias rating risk of 'high' for any study characteristic as determined by Memtsoudis *et al* using the Cochrane Risk of Bias Tool were included in the high risk of bias subgroup; all other studies were treated as having a low risk of bias.

The specific division for the subgroups divided by patients per study was studies with greater than the median number of patients of all the studies analyzed versus those that did not.

Sentiment analysis

The algorithm used for sentiment analysis in this study, GAN-BioBERT, was written based on a semisupervised version of the previously described bidirectional encoder representations from transformers (BERT) algorithm for natural language processing by Devlin *et al*.[7 8] GAN-BioBERT was built by Myszewski *et al* specifically for categorically classifying the tone expressed in clinical abstracts as either positive, negative, or neutral. Examples of text classified into each of these categories are shown in table 1.

To understand the utility of GAN-BioBERT as it relates to this study, we first need to discuss its precursors, BERT and BioBERT. The process by which GAN-BioBERT comes from these previous methods is shown graphically in figure 1.

A detailed description of the original BERT algorithm is available elsewhere,[7] but some discussion of the algorithm and its derivations that led to the algorithm used in this study is important to understanding its applicability to clinical literature.

The original BERT algorithm is currently considered as one of the state-of-the-art methodologies for natural language processing with high levels of accuracy. This algorithm uses a methodology known as transfer learning wherein a pretrained language model for a particular domain (ie, biomedical literature) is first developed with an extremely large sample of text from the language domain being studied. The original BERT model by Devlin *et al* was trained on a set of 2.5 billion words of text from English Wikipedia as well as 800 million words of text from BooksCorpus, a large collection of English language novels.[7] This original BERT model was designed to be suited for general language tasks, but not to understand the nuances and complicated language frequently used in biomedical literature.

This general model was then refined in a second 'fine-tuning' step with fewer samples for the task of interest, (ie, biomedical literature) in a method that is dependent on the appropriateness of the original general language model.

To make this general language model more applicable to biomedical literature BioBERT was developed in 2020 by Lee *et al*.[9] This was done by further training the original BERT model with an additional 4.5 billion words from PubMed abstracts as well as 13.5 billion words from PubMed Central full-text articles to create the language model known as BioBERT,[9] which is uniquely designed/appropriate for the nuances of the writing style and terminology used in biomedical and academic literature

The algorithm/language model used in this study, GAN-BioBERT, was developed by Myszewski *et al* by further fine-tuning the biomedically oriented language model BioBERT for the specific task of classifying clinical study abstract sentiment.

The term fine-tuning is used to describe the process of refining a pretrained language model, that is, BioBERT, for a particular task such as sentiment classification of clinical trial abstracts by providing a smaller task specific set of examples to the algorithm. Following this fine-tuning step, the algorithm's performance is assessed and can then be applied for the proposed application (ie, sentiment classification of biomedical abstracts).

The sentiment classifications made by the GAN-BioBERT algorithm used for this study coincided with the determination of clinicians 91.3% of the time for classifying the sentiment in clinical trial abstracts as positive, negative, or neutral. This was determined by comparing the sentiment classifications made by the algorithm to the categorical classifications made by a set of clinicians on a set of sample abstracts that contained an equal

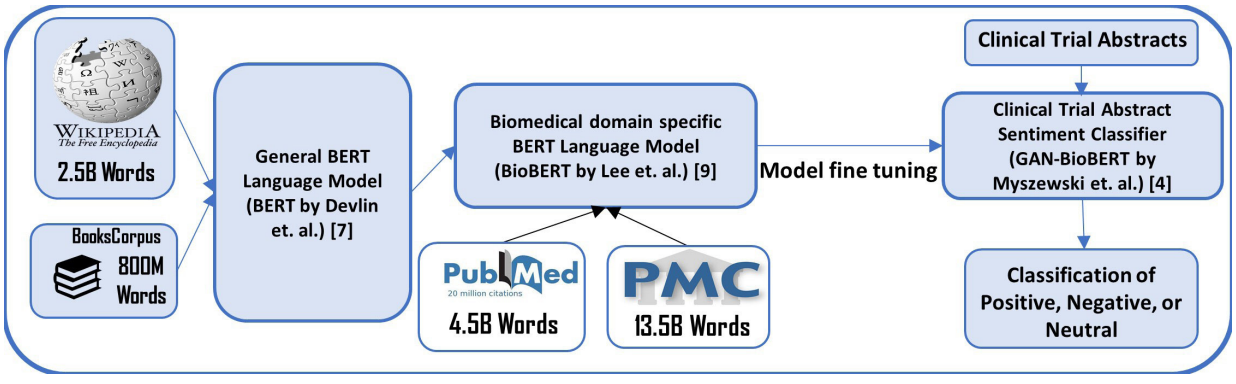

Figure 1 A graphical representation of the development of GAN-BioBERT as it relates to this study. BERT, bidirectional encoder representations from transformers.

amount of positive, negative, and neutral abstracts as determined by the clinicians.[4]

It is important to note that on an individual article level, GAN-BioBERT only categories abstracts as positive, negative, or neutral. This three-category classification scheme was chosen to provide adequate detail without sacrificing an adequate level of accuracy. More finely grained algorithms for sentiment classification are significantly limited by drops in accuracy as the granularity of the sentiment classifications is increased. For example, the original BERT algorithm was correct only 55.5% of the time when used for a sentiment classifier with five categories, as compared with being correct 93.1% of the time for a two-category benchmark dataset.[7] Given this limitation, GAN-BioBERT is only suited to large samples of articles and not for use on individual articles.

RESULTS

Of the 122 study abstracts included in the original meta-analysis, 115 were included for sentiment analysis. One article was excluded due to not having an abstract but was still indexed in PubMed and six other articles were excluded due to not being indexed within the PubMed database. Each study, alongside its corresponding sentiment value determined by the algorithm is provided in online supplemental appendix A1.

Of the 115 articles included for analysis, it was determined that 56 (46.1%) had positive sentiment, 22 (19.1%) had negative sentiment, and 50 (34.8%) were neutral using sentiment analysis. The entire runtime for the analysis program, including data gathering, sentiment classification, and the subgroup analysis, was 8 min and 39 s on a desktop computer.

For comparison, the findings of the meta-analysis by Memtsoudis *et al* expressed a generally positive outlook toward the application of peripheral nerve block analgesia for total hip and knee arthroplasty.[5] This sentiment was determined qualitatively based on the general recommendation made in that study that the use of peripheral nerve block analgesia is recommended for hip and knee arthroplasty as it leads to improved clinical outcomes. This generally positive outlook in the meta-analysis is aligned with the findings determined using sentiment analysis where the greatest proportion of studies had positive sentiment. It is also important to note that a large amount of the studies included expressed neutral sentiment, indicating a more moderate positive outlook.

The results for each of the subgroup analyses are shown in table 2. The median number of patients per study was found to be 80 patients.

DISCUSSION

In this study, a sentiment analysis of the sources used by a recent meta-analysis evaluating the clinical efficacy of regional anesthesia procedures for hip and knee arthroplasty reached a similar conclusion to that of the meta-analysis. In doing so, this current study accomplished two important goals that warrant further discussion. First, this study demonstrated that the GAN-BioBERT sentiment analysis approach yielded results concurrent with the findings of the Memtsoudis *et al* meta-analysis. Second, this study demonstrated that sentiment analysis represents an efficient alternative and/or complement to meta-analysis studies, with the total runtime of the analysis program requiring fewer than 10 min.

This study's (and the GAN-BioBERT algorithm's) major limitation is created by the three-class classification system and manifests as an inability to draw conclusions without an adequate sample size. This limitation restricts the use of the algorithm such that it cannot be used to draw conclusions on an individual level and is most appropriately used on a large aggregative scale to identify publication trends or as a precursor to more in-depth assessment of academic literature. With this in mind, the algorithm's application is limited to assessing larger trends such as topic-specific sentiment, as well as how this sentiment varies across time or specialty. Additional examples of possible applications fitting with this limitation include identifying trends related to publication biases as was shown in the subgroup analysis where the high bias and low bias subgroups had significantly different findings. Ideally, as the technology of sentiment analysis continues to advance, more granular classification schemes will achieve acceptable accuracy levels and the large sample size requirement will be ameliorated.

This study does possess several other limitations that should be considered prior to widespread application of sentiment scoring for guiding clinical decision-making. First, this specific algorithm only examines vocabulary located within abstracts of published studies. This limitation is imposed secondary to a significant loss of classification accuracy when the technology is used to classify sentiment for longer length bodies of text such as those in the body of manuscripts.[1] Therefore, text within the body of the manuscript that may recommend tempered

Table 2 Proportion of positive, negative, and neutral abstracts in each subgroup

Subgroup	Sample (n)*	Positive	Negative	Neutral
All included studies	115	53 (46.1%)	22 (19.1%)	50 (34.8%)
Both general and neuraxial anesthesia	26	12 (46.2%)	7 (26.9%)	7 (26.9%)
Only general anesthesia	24	12 (50%)	7 (29.2%)	5 (20.8%)
Only neuraxial anesthesia	49	22 (44.9%)	5 (10.2%)	22 (44.9%)
Hip arthroplasty	27	12 (44.4%)	5 (18.5%)	10 (37.1%)
Knee arthroplasty	92	41 (44.6%)	19 (20.6%)	32 (34.8%)
Studies with >80 patients	55	17 (30.9%)	17 (30.9%)	21 (38.2%)
Studies with <80 patients	60	36 (60%)	5 (8%)	19 (32%)
High risk of bias studies	37	23 (62.2%)	3 (8.1%)	11 (29.7%)
Low risk of bias studies	78	30 (38.5%)	19 (24.4%)	29 (37.1%)
Observational studies	40	15 (37.5%)	11 (27.5%)	14 (35%)
Randomized control trials	75	38 (50.7%)	11 (14.7%)	26 (34.6%)

*These subgroups are not all mutually exclusive and there may be overlap between groups.

enthusiasm or shades of optimism would not be subject to evaluation via the reported evaluation method. Second, this study only used sentiment analysis as a mechanism to validate a single meta-analysis study. It is possible that further examination of diverse meta-analysis studies may not result in similar findings or may find that different specialties are more apt to use positive, negative, or neutral terminology within their scientific writing. Third, the use of sentiment analysis in this context is unable to determine the impact of the studied intervention on a variety of clinical outcome domains. For example, the study by Memtsoudis *et al* demonstrated that the use of peripheral nerve blocks in the setting of knee and hip arthroplasty reduced the risk of cognitive dysfunction, respiratory failure, cardiac complications, and surgical site infections. However, in some cases, sentiment analysis may align closer to clinical decision-making where a clinician plans to administer a 'better' option and not necessarily one that offers an improvement in any distinct outcome domain.

Finally, performing a sentiment analysis may be beyond the technology limitations of an individual and therefore may not represent a feasible mechanism of literature review for all providers. Furthermore, while this study showed that sentiment analysis could accelerate the process of assessing literature, the application of sentiment analysis will still require the same careful consideration of the sources used for a particular research question or topic as is currently used in the process of systematic analyses and meta-analyses. However, it is important to consider that meta-analysis and expert opinion are imperfect and potentially subject to biases as rules are variably applied to which manuscripts are included in the final analysis or opinion is based on local experience and training.

Briefly, this study successfully used sentiment analysis as a rapid and efficacious mechanism to reach the same conclusion as a previously published meta-analysis study. With this in mind, sentiment analysis shows promise as a clinical literature evaluation tool that can be added to the repertoire of methods researchers and clinicians use to perform their work, verify study findings, and generate novel hypotheses.

Twitter Kristopher M Schroeder @KristopherSchr6

Contributors JM performed analysis, data gathering, manuscript writing, and project conception. EK performed manuscript editing and writing, analysis, and interpretation of the results. KMS provided supervision, guidance and institutional support, performed manuscript writing, and assisted with project conception.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD
Joshua Myszewski http://orcid.org/0000-0002-3690-7056

REFERENCES

1 Choi Y, Lee H. Data properties and the performance of sentiment classification for electronic commerce applications. *Inf Syst Front* 2017;19:993–1012.
2 Ceron A, Curini L, Iacus SM, *et al*. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 2014;16:340–58.
3 Xu J, Zhang Y, Wu Y, *et al*. Citation sentiment analysis in clinical trial papers. *AMIA Annu Symp Proc* 2015;2015:1334–41.
4 Myszewski JJ, Klossowski E, Meyer P. Validating GAN-BioBERT: a methodology for assessing reporting trends in clinical trials. *arXiv preprint arXiv* 2021:2106.00665.
5 Memtsoudis SG, Cozowicz C, Bekeris J, *et al*. Peripheral nerve block anesthesia/analgesia for patients undergoing primary hip and knee arthroplasty: recommendations from the International consensus on Anesthesia-Related outcomes after surgery (ICAROS) group based on a systematic review and meta-analysis of current literature. *Reg Anesth Pain Med* 2021;46:971–85.
6 Myszewski J. jmyszewski/GANBioBERTMetaAnalysis: initial release v1.0. *GitHub release* 2021.
7 Devlin J, Chang MW, Lee K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv* 2018:1810.04805.
8 Croce D, Castellucci G, Basili R. GAN-BERT: generative adversarial learning for robust text classification with a bunch of labeled examples. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020:2114–9.
9 Lee J, Yoon W, Kim S, *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36:1234–40.