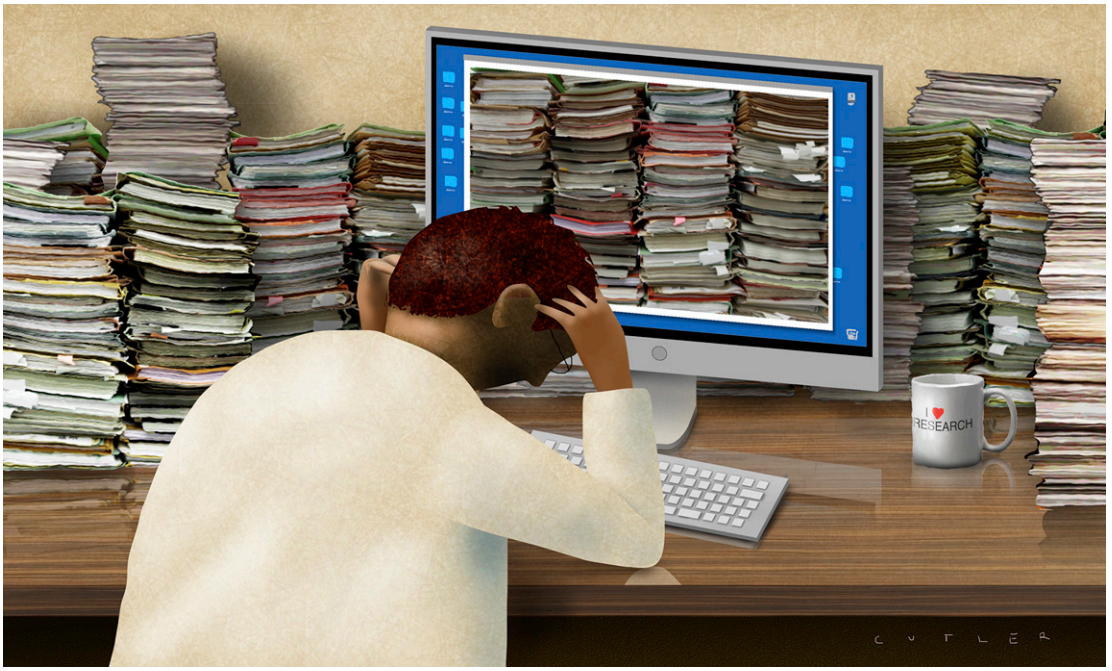
Medical misinformation in the era of Google: Computational approaches to a pervasive problem

Scott R. Granter[a,b] and David J. Papke Jr.[a,1]

On December 28, 1917, a fascinating article appeared in the pages of the *New York Evening Mail*. The article, titled "A Neglected History," written by H.L. Mencken, laments the fact that the 75th anniversary of the introduction of the bathtub to the United States had passed without the slightest public notice. "Not a plumber fired a salute or hung out a flag. Not a governor proclaimed a day of prayer. Not a newspaper called attention to the day" (1). Mencken goes on to detail the history of the bathtub, describing the introduction of the English bathtub by Lord John Russell in 1828, "then, as now … a puny and inconvenient contrivance—little more, in fact, than a glorified dishpan." Mencken wrote that installation of the bathtub in Millard Fillmore's White House in 1851 led to more widespread acceptance in the United States. Quoting from the purported April 23, 1843, issue of *Western Medical Repository*, he goes on to record the opposition of physicians to the bathtub as dangerous to health, inviting "phthisic, rheumatic fevers, inflammation of the lungs, and the whole category of zygomatic diseases." And he states that by 1859 the majority of the medical community had finally accepted the bathtub as harmless to health—evidenced by a poll taken at the 1859 meeting of American Medical Association in Boston, in which nearly 55% of physicians regarded the bathtub as harmless and more than 20% advocated its use as beneficial to health (1).

The story was widely read, generated much interest, and was widely quoted. The problem is very little in the story was true. Most of the facts presented



The spread of misinformation in science and medicine is a real problem, not only in spite of technology but often because of it. To sort through the cavalcade of journal articles, legitimate and otherwise, the scientific community should devote more resources to technological approaches that identify false and retracted findings. Image courtesy of Dave Cutler (artist).

[a]Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115; and [b]Department of Pathology, Harvard Medical School, Boston, MA 02115

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: dpapke@partners.org.

were "absurdities, all of them deliberate and most of them obvious" (2). Mencken finally, on May 23,1926, in the pages of the *Chicago Tribune*, confessed and informed the reading public of his ruse, which he said was meant as a diversion, an attempt at "levity" during the strain of war (2). By the time he published his confession, Mencken was disturbed to see his "preposterous facts ... cited by medical men ... in learned journals" (2). But Mencken was very much mistaken if he thought coming clean would finally put to rest the false history he had created. Mencken's fabricated story continued to be cited as fact for decades after his correction of the record. The horse was out of the barn and could not be led back in.

The 21st-century reader, with smartphones and Google at the ready, can reflect on the readers of Mencken's original essay with sincere sympathy. The cumbersome nature of fact checking in 1917 would likely prevent the average newspaper reader from researching the truth of the matter. Some modern readers might also believe that contemporary biomedical literature, with rigorous peer review and ample opportunity to respond to and comment on papers, would not be subject to persistence of misinformation (using an inclusive definition of misinformation encompassing raw "information" as well as synthesized "knowledge"). However, one study found that a set of widely publicized fraudulent articles that were retracted in 1981 continued to be cited 298 times in the English-language literature from 1982 to 1990 (3). Of the citing articles, 85.9% cited the retracted work positively, 8.4% cited it negatively, and only 5.7% of citations acknowledged the fraudulent nature of the retracted articles.

It is understandable that in the pre-Internet era retracted articles would persist—after all, you can't retract a bound volume in a library or a reprint in a personal office. But the problem of retracted articles has not gone away with the rise of Internet databases. Studies have shown that, even in the era of Google, articles continue to accrue hundreds of citations after retraction (4, 5). A subset of retracted articles continued to appear in centralized online databases as late as 2011 without indication of retraction (6), and many retracted articles can be found on non-publisher websites (7). The authors of the latter study concluded "the benefits of decentralized access to scientific papers may come with the cost of promoting incorrect, invalid, or untrustworthy science" (7). Misinformation is still a problem today, not only in spite of technology but often because of it. Here, we propose that the scientific community devotes more resources to technological approaches that address misinformation.

Many Facets of Misinformation

Unfortunately, the citation of retracted articles is only the tip of the iceberg of biomedical misinformation. In a 1947 address, Charles Sidney Burwell, the 18th dean of Harvard Medical School, acknowledged that "Your teachers have tried to give you a good opportunity to learn and to offer you information which the evidence indicated to be accurate. Nevertheless, probably half of what you know is no longer true. This troubles me, but what troubles me more is I don't know which half it is" (8). It turns out that not much has changed since Burwell's day. Richard Horton, current editor-in-chief of *The Lancet*, echoes Burwell: "The case against science is straightforward: much of the scientific literature, perhaps half, may simply be untrue" (9). Marcia Angell, a former editor-in-chief of the *New England Journal of Medicine* said, "it is simply no longer possible to believe much of the clinical research that is published" (10). Biomedical misinformation is a much broader problem than the inappropriate citation of retracted articles.

There are strong statistical arguments that most scientific claims in peer-reviewed journals are false (11). The prevalence of false claims would not be a problem on its own if these works were discounted in the face of stronger, contradictory evidence. We are often told that science is, after all, characterized by "self-correction." However, Tatsioni et al. (12) showed that relatively weak observational studies are cited favorably despite publication of stronger randomized control studies providing evidence to the contrary, with around half the citations of the weaker study remaining favorable. For example, the authors found that articles continued to refer to the cardiovascular benefits of vitamin E even after these results could not be replicated by two randomized trials and that 61.7% of articles favorably cited a protective effect for estrogen on Alzheimer's disease long after this claim had been strongly contradicted (12).

The prevalence of false claims would not be a problem on its own if these works were discounted in the face of stronger, contradictory evidence.

There is a range of interrelated issues that lead to such trends—on one end of the spectrum, articles retracted for fraud or critical errors continue to be inappropriately cited, akin to H.L. Mencken's "A Neglected History," and on the other end, weak but arguably valid evidence continues to be cited without mention of stronger contradictory evidence. The full spectrum, ranging from fraud to innocent statistical errors to unqualified citation of strongly contradicted studies, is unified by two characteristics: (a) There exist technological approaches to mitigate the propagation of all these types of misinformation, and (b) the scientific community places little emphasis on addressing these problems, as reflected by the many-thousandfold difference in funding data organization compared with the funding for data generation.

Computational Approaches

We currently have the ability to extirpate the unmitigated citation of retracted articles—although not trivial, the creation and maintenance of a centralized list of retracted publications would be inexpensive relative to the amount of money being funneled into

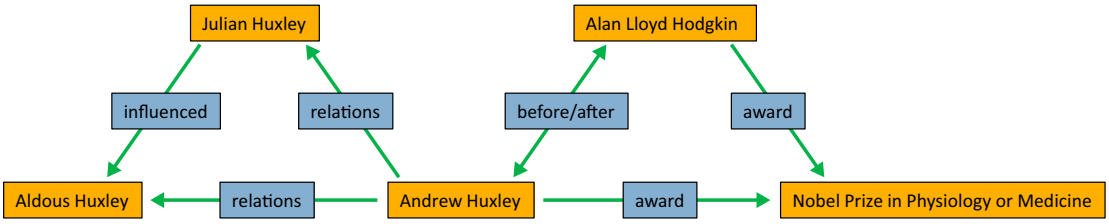


Fig. 1. This partial knowledge graph demonstrates the connectivity between Aldous Huxley, author of *Brave New World* (17), and the Nobel Prize in Physiology or Medicine, which was awarded to his brother Andrew Huxley in 1963 for research on ionic currents across membranes. Concept nodes are colored orange and connection labels are colored green. Andrew Huxley and the Nobel Prize are directly connected by an edge in the graph (labeled "award"), while Aldous Huxley and the Nobel Prize are not directly connected by an edge. Graph generated in part using RelFinder (18).

basic science research. Work to this end is already underway: In 2016 the MacArthur Foundation allocated $400,000 to the Center for Scientific Integrity's Retraction Watch to establish a comprehensive database of retracted articles (13). Once this database is in place, journals can easily implement an algorithm to cross-check all citations in papers under review against the database and to automatically flag citations of retracted articles to call them to the attention of reviewers. Such an approach would ensure that the scientific community does not continue to cite retracted articles to inappropriately support further research.

It is more difficult to address the citation of weak evidence without mention of stronger contradictory evidence. This problem is at least in part attributable to the labor-intensive, nonlinear nature of literature searches—for example, a PubMed search for "Alzheimer's disease" and "estrogen" yields a dauntingly long list of articles from which it is difficult to quickly assess the state of evidence supporting a potential connection. It would be more useful to generate output organized by level of evidence from which the state of the field could be more easily gleaned. There has been some work to summarize and synthesize knowledge from disparate sources in the literature, most notably via Cochrane and the Campbell Collaboration for systematic reviews and meta-analyses. However, as evidenced by aforementioned studies, there is significant room for improvement, and Campbell Collaboration and Cochrane researchers have called for development of novel technological approaches to synthesizing knowledge (14).

The path forward is clear: The scholarly community should use technological approaches to maintain the integrity of the published literature.

There has been a movement in computer science to generate so-called "knowledge networks" for automated fact checking by using online databases of facts. "DBpedia," perhaps the best-known example, is a large-scale community effort to extract data from Wikipedia infoboxes (the boxes in Wikipedia that contain purely declarative data, such as the population of a city or the names of members of a musical group). Data are automatically extracted from Wikipedia by algorithms and are entered into the DBpedia database, which is optimized to be easily parsed by algorithms (15). For example, the fact that Andrew Huxley won the Nobel Prize is listed in Huxley's Wikipedia infobox, so these data appear in the DBpedia database as ["Andrew_Huxley," "award," "Nobel_Prize_in_Physiology_or_Medicine"], a format that is easily queried by algorithms.

In recent work by Ciampaglia et al. (16), this DBpedia database was used to construct a "Wikipedia Knowledge Graph" in which the people, places, and things in infoboxes—collectively known as "concept nodes"—were represented as vertices in the graph, and connections, or "edges," were drawn between these vertices if concept nodes were connected by the appropriate type of statement (Fig. 1). The authors were interested in using this graph to assess the accuracy of statements not directly represented in the graph. For example, the content in the statement "Aldous Huxley's brother won the Nobel Prize" would not be directly represented in the graph, but this connection is indirectly contained in the graph via links between ["Aldous_Huxley," "relations," "Andrew_Huxley"] and ["Andrew_Huxley," "award," "Nobel_Prize_in_Physiology_or_Medicine"] (Fig. 1). For indirect statements regarding presidential spouses, the algorithm was 98% successful in judging truthfulness. More importantly for scientific applications, it would be easy to test the veracity of statements that are directly represented in the graph.

A limitation to translating this work to the scientific literature is the need for well-defined input parameters for data gleaning. Unfortunately, scientific articles do not have infoboxes with distilled conclusions, which raises the question of how we might be able to generate summary statements from articles and whether this summarization can be automated. There have been recent advances in correctly classifying text into categories; for example, algorithms have been developed that can correctly classify Amazon reviews as being positive (four or five stars out of five) or negative (one or two stars out of five) with an accuracy of 94.49%, despite the algorithm having no prior knowledge of word definitions (19). However, the field still has a long way to go—state-of-the-art algorithms still cannot accurately choose the correct sentence to end a short story (20). Some speculate that text

summarization will require truly artificially intelligent machines (21).

One currently feasible alternative is for humans to be involved in the initial step of distilling a few simple conclusions from articles. These extracted data can then be fed into algorithms to organize and highlight inconsistencies across the literature. The above-outlined knowledge graphs are an avenue for exploration in this context. With minimal effort, authors, reviewers, or editors could submit a few declarative statements describing conclusions of an article—such as "estrogen benefits Alzheimer's patients" or "vitamin E does not have cardiovascular benefits"—to enable construction of a knowledge graph similar to the Wikipedia Knowledge Graph (16). Authors are already asked to submit keywords and brief summaries along with articles, and submission forms could easily be modified to include a field for summary sentences. Furthermore, the connections in the resulting graph could be labeled by level of evidence, so that for any given relationship (i.e., Alzheimer's and estrogen) there would be a list of supporting and contradictory evidence, categorized by level of support. Such an approach would lead to better organization at the inherent expense of oversimplification. However, as in biology, there is utility in taking both reductionist and holistic approaches. Insights gained from a holistic overview of the literature would be useful in combatting erroneous citations.

Such approaches would require some effort from the scientific community, as well as a paradigm shift toward increased emphasis on data curation. A Freedom of Information Act request revealed that the 2013 PubMed Central budget was $4.45 million (22). In contrast, the budget of the NIH alone was $29.15 billion in the same year (23), $14.9 billion of which was granted for research projects (24). Effectively, our funding system puts several thousand times more emphasis on data generation than it does on data organization. As the number of publications continues to increase, and along with them the number of erroneous citations and contradicted publications, we foresee the need to shift more focus to organization.

As of this writing, a PubMed search returns more than 28 million shambolic biomedical citations. The persistence of misinformation in the literature is perplexing in the era of Google. We have the ability now to prevent the unqualified citation of retracted articles, and computational avenues could help curtail the unqualified citation of studies that have been contradicted by studies possessing a higher degree of evidence. The path forward is clear: The scholarly community should use technological approaches to maintain the integrity of the published literature.

Acknowledgments

We thank Yulia Maximenko and Codruta Girlea for useful discussions.

1 Mencken HL (December 28, 1916) A neglected anniversary. *New York Evening Mail*.
2 Mencken HL (May 23, 1926) Melancholy reflections. *Chicago Tribune*, Part 8, p 2.
3 Kochan CA, Budd JM (1992) The persistence of fraud in the literature: The Darsee case. *J Am Soc Inf Sci* 43:488–493.
4 Unger K, Couzin J (2006) Scientific misconduct. Even retracted papers endure. *Science* 312:40–41.
5 Neale AV, Northrup J, Dailey R, Marks E, Abrams J (2007) Correction and use of biomedical literature affected by scientific misconduct. *Sci Eng Ethics* 13:5–24.
6 Wright K, McDaid C (2011) Reporting of article retractions in bibliographic databases and online journals. *J Med Libr Assoc* 99:164–167.
7 Davis PM (2012) The persistence of error: A study of retracted articles on the Internet and in personal libraries. *J Med Libr Assoc* 100:184–189.
8 Burwell CS (1947) The medical school. *Harv Med Alumni Bull*, October, p 6.
9 Horton R (2015) Offline: What is medicine's 5 sigma? *Lancet* 385:1380.
10 Angell M (2009) Drug companies and doctors: A story of corruption. *New York Rev Books* 56:8–12.
11 Ioannidis JPA (2005) Why most published research findings are false. *PLoS Med* 2:e124.
12 Tatsioni A, Bonitsis NG, Ioannidis JPA (2007) Persistence of contradicted claims in the literature. *JAMA* 298:2517–2526.
13 Kretser A, Murphy D, Dwyer J (2017) Scientific integrity resource guide: Efforts by federal agencies, foundations, nonprofit organizations, professional societies, and academia in the United States. *Crit Rev Food Sci Nutr* 57:163–180.
14 Elliott JH, et al. (2015) Informatics: Make sense of health data. *Nature* 527:31–32.
15 Lehmann J, et al. (2015) DBpedia: A large-scale, multilingual knowledge base extracted from Wikipedia. *Semant Web* 6:167–195.
16 Ciampaglia GL, et al. (2015) Computational fact checking from knowledge networks. *PLoS One* 10:e0128193.
17 Huxley A (1932) *Brave New World* (Chatto & Windus, London).
18 Heim P, Hellmann S, Lehmann J, Lohmann S, Stegemann T (2009) RelFinder: Revealing relationships in RDF knowledge bases. Proceedings of the Fourth International Conference on Semantic and Digital Media Technologies (SAMT 2009), eds Chua TS, et al. (Springer, Berlin), pp 182–187.
19 Zhang X, LeCun Y (2015) Text understanding from scratch. arXiv.1502.01710.
20 Mostafazadeh N, et al. (2016) A corpus and cloze evaluation for deeper understanding of commonsense stories. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Available at www.aclweb.org/anthology/N16-1098. Accessed May 15, 2018.
21 Ide N, Veronis J (1998) Introduction to the special issue on word sense disambiguation: The state of the art. *Comput Linguist* 24:2–40.
22 Anderson K (2013) The price of posting—PubMed Central spends most of its budget handling author manuscripts. *The Scholarly Kitchen*. Available at https://scholarlykitchen.sspnet.org/2013/07/16/the-price-of-posting-pubmed-central-spends-most-of-its-budget-handling-author-manuscripts/. Accessed February 21, 2017.
23 Rockey S (2013) Funding operations for FY2013. National Institutes of Health, Office of Extramural Research. Available at https://nexus.od.nih.gov/all/2013/05/08/funding-operations-for-fy2013/. Accessed February 21, 2017.
24 Rockey S (2014) FY2013 by the numbers: Research applications, funding, and awards. National Institutes of Health, Office of Extramural Research. Available at https://nexus.od.nih.gov/all/2014/01/10/fy2013-by-the-numbers/. Accessed February 28, 2017.