Nuggets: findings shared in multiple clinical case reports

Neil R. Smalheiser, MD, PhD; Weixiang Shao, MS; Philip S. Yu, PhD

See end of article for authors' affiliations

DOI: http://dx.doi.org/10.3163/1536-5050.103.4.002

Objective: The researchers assessed prevalence in the clinical case report literature of multiple reports independently reporting the same (or nearly the same) main finding.

Methods: Results from forty-five PubMed queries were examined for incidence and features of main findings ("nuggets") shared in at least four case reports.

Results: The authors found that nuggets are surprisingly prevalent and large in the case report literature, the largest found so far was reported in seventeen articles. In most cases, the main findings of case reports were evident from examining titles alone.

Conclusions: Our curated examples should serve as gold standards for developing specific automated methods for finding nuggets. Nuggets potentially enable finding-based (instead of topic-based) information retrieval.

Keywords: Evidence-Based Medicine, Medical Informatics, Information Storage and Retrieval, Case Reports

INTRODUCTION

The problem of identifying subsets of documents that "say (more or less) the same thing" is related to, but distinct from, existing text-mining techniques that seek to classify or cluster documents according to overall similarity of the topics that they discuss, to identify predominant themes within the literature, or to extract or summarize knowledge from an entire body of documents [1]. Clinical case reports are an ideal type of biomedical literature for undertaking this type of analysis, for medical as well as methodological reasons. Roughly 1.7 million articles are indexed in PubMed as case reports, about 7% of all biomedical articles; yet case reports rank at the bottom in the hierarchy of evidence-based medicine, far below randomized controlled trials. Case reports are generally not included in the assessment of clinical evidence carried out by systematic reviews and meta-analyses. Case reports are generally poorly cited relative to other research articles [2, 3], leading some journals either to stop accepting case reports or to classify them as "letters" or "comments" that will not affect the journal's impact factor. Like any other type of eyewitness reports, case reports may be prone to observer bias, may reflect wrong assumptions or premature interpretation of findings, and generally lack controlled conditions so that, at best, findings are correlative instead of conclusive.

On the other hand, eyewitness reports often provide the first observations of new phenomena or new innovations. For example, three case reports appearing in 1981 provided the first delineation of AIDS [4–6]. Randomized clinical trials generally follow patients for short periods and may be subject to sponsor bias, so case reports may be a more reliable, unbiased venue for reporting long-term adverse effects. Recent editorials have argued that case reports provide essential evidence in medicine [7]. In fact, there has been a recent renaissance of the case report literature. Leading journals such as *Lancet* continue to publish case reports regularly, and scores of new journals devoted specifically to case reports have emerged from publishers such as BMJ, Elsevier, Wiley, Oxford University Press, and Hindawi. Case

This article has been approved for the Medical Library Association's Independent Reading Program <http://www.mlanet.org/education/irp/>.

Supplemental Table 2 and Table 3 are available with the online version of this journal.

reports may not be entirely independent of each other (since publishing one report may spur interest in reporting additional patients), but except in rare situations [8], each article may be taken as an independent source of evidence.

Methodologically, case reports are a favorable test bed for exploring discovery of main findings because each case report tends to be short and concise. Case reports tend to state a single main finding that is often directly stated in its title. This situation is much simpler than occurs for some publication types, such as clinical trials or genetic association studies [9].

In summary, case reports are a valuable, unique, yet noisy and underutilized type of evidence. The authors believe that there is considerable value in identifying findings that have been independently published in multiple case reports, since that would alert readers to evidence that has particularly high reliability and potential impact. In turn, this might encourage wider judicious use of case reports in evidence-based medicine and other tasks such as surveillance of drug side-effects [10]. Despite the fact that many case report articles are accessible and indexed in MEDLINE, no automated tool exists that can find subsets of articles that report the same or very similar main findings.

METHODS

Search strategies

Overall, about forty-five PubMed queries were performed and analyzed to learn whether there were, indeed, multiple case reports that stated the same or almost the same main finding. We also attempted to validate our expectation that the main finding of a case report could generally be discerned by simply looking at the title.

First, in an exploratory strategy, fifteen ad hoc PubMed queries were performed in July 2011 by entering the name of a common drug *and* the name of a common disease (limited to case reports [Publication Type]) into Anne O'Tate, a value-added PubMed search engine that is publicly accessible [11, 12]. Anne O'Tate offers a "cluster by topic" option that allows the user to divide the retrieved articles into no more than a dozen distinct topics using Medical Subject Heading (MeSH) terms, which facilitates examination of relatively small, homogeneous subsets that share common topics. Anne O'Tate also offers an "important words" option that displays a list of title and abstract words ranked according to their document frequency in the search output, relative to the overall frequency in MEDLINE as a whole [11].

In a second, more systematic strategy, we tabulated all MeSH terms that were relatively infrequent (i.e., that were indexed within a total of fifteen to thirty-five articles in MEDLINE). We randomly selected thirty of these MeSH terms and formulated PubMed queries of the form "MeSH term"[MeSH] AND case reports[Publication Type]. The titles (and in selected cases, abstracts or full text) of the retrieved articles were manually examined by Smalheiser and scored on two separate occasions more than six months apart to ensure test-retest validity.

Nuggets

A "nugget" is defined as a main finding reported in four or more case reports (within the set of articles retrieved by a given PubMed query, which will be referred to as the query set). In general, there may be no, one, or more than one nugget within a set of case reports. The threshold value of four articles is arbitrary but was chosen to exclude situations such as duplicate publication of the same article in different languages or in different journals. The main finding must be the motivating reason for publishing the report, must be highly similar or identical across the case reports, and must not merely refer to the same problem or topic. In general, a main finding might not be discernible simply from reading the titles of the case reports; however, the present study only attempts to identify nuggets that can be discerned from titles. The set of articles that share the same main finding are referred to as the articles that map to the nugget. Note that the number of articles that map to the nugget is largely independent of the size of the query set; thus, a nugget may not be a prevalent finding or predominant theme of the query set as a whole.

Graph-based characterization

The titles were tokenized (i.e., split into distinct words), and stemmed and lemmatized, which reduce inflected word forms to base forms (e.g., "walking" and "walked" are both converted to "walk") using the Stanford CoreNLP package [13]. These term-processing steps allowed us to recognize words as shared across different titles despite minor differences such as singular versus plural or present versus past tense. The tokenization scheme was

Zolpidem Dependence and Withdrawal Seizure—Report of Two Cases”
“***Zolpidem Dependence and Withdrawal Seizure***”
“A Fatal Case of Benzodiazepine Withdrawal”
“Intoxication with a Tricyclic Antidepressant”
“*Dependence on Zolpidem*”
“**Seizure Following Sudden Zolpidem Withdrawal**”
“***Zolpidem at Supratherapeutic Doses Can Cause Drug Abuse, Dependence and Withdrawal Seizure***”
“*Dependence on Zolpidem: A Report of Two Cases*”
“**Zolpidem-Related Epileptic Seizures: A Case Report**”
“**Seizures Associated with Venlafaxine, Methylphenidate, and Zolpidem**”
“**Epileptic Seizures as a Sign of Abstinence from Chronic Consumption of Zolpidem**”
“***Abuse, Dependence, and Epileptic Seizures after Zolpidem Withdrawal: Review and Case Report***”
“Radiopacity of Clomipramine Conglomerations and Unsuccessful Endoscopy: Report of 4 Cases”
“**Seizure after Withdrawal from Supratherapeutic Doses of Zolpidem Tartrate, a Selective Omega I Benzodiazepine Receptor Agonist**”
“Acute Poisoning by New Psychotropic Drugs”
“*Chronic Abuse of Zolpidem*”

“Sixteen articles were retrieved by the PubMed query carried out in July 2011: [(zolpidem OR ambien OR edluar) AND (seizure OR seizures) AND (case reports[PT] OR (english abstract[PT] AND case[TIAB])). A simple inspection of titles (listed in reverse chronological order) reveals that at least six articles (in italics) report the potential of zolpidem for abuse and/or dependence, and at least nine articles (in bold) report the occurrence of seizures related to zolpidem (of which seven explicitly relate this to acute withdrawal or abstinence from zolpidem).

Table 1

Titles retrieved from zolpidem and seizures query

modified to capture certain term variants as identical (e.g., DARPP32 would be merged with DARPP 32 and DARPP-32) and to count hyphenated words as comprising 3 separate words (e.g., A-B would be counted as three tokens, A-B, A, and B, though we would not count links among the 3 when constructing cliques). A stoplist of 768 frequent words was taken from an online stoplist created by Damian Doyle [14] that we modified to include single letters and the specific medical terms “case” and “report.” No attempt was made to collapse synonyms or lexical variants (e.g., American vs. British spelling) nor to recognize multiword terms as biologically unitary entities. Nodes were defined as processed title words, and links were defined as joining two nodes that co-occur in the same title. Link weight is the number of articles in which the two nodes co-occur. Cliques are defined as any set of nodes where each node is connected to all others, satisfying minimal criteria (node document frequency >3, node number >3, and link weight >2). Note that a clique may be present that has, for example, 5 nodes, even if no single title mentions all 5 nodes. In that sense, cliques represent knowledge across a set of documents.

RESULTS

Analysis of ad hoc PubMed queries

At the outset, it was not obvious whether nuggets would be prevalent at all. However, it was surprisingly easy to find main findings reported by 4 or more case reports. For example, beginning with one Anne O’Tate query, [(seizure OR seizures) AND psychiatric AND (case reports[pt] OR (english abstract[pt] AND case[tiab]))], that retrieved 803 articles, we drilled down to examine articles that mentioned certain “important words” restricted to clinical drug or pharmacologic substance semantic categories [11]. Zolpidem (trade names include Ambien®, Edluar®, Intermezzo®, and Zolpimist®) was one of several agents that were mentioned in multiple case reports. To examine zolpidem in detail, a new focused query was performed: [(zolpidem OR ambien OR edluar) AND (seizure OR seizures) AND (case reports[PT] OR (english abstract[PT] AND case[TIAB]))]. As shown in Table 1, a simple inspection of titles revealed that at least 6 articles (in italics) reported the potential of zolpidem for abuse and/or dependence, and at least 9 articles (in bold) reported the occurrence of seizures related to zolpidem, of which 7 explicitly related this to abuse and/or acute withdrawal or abstinence from zolpidem. The other articles in this list had titles that were relatively uninformative (“A Fatal Case of Benzodiazepine Withdrawal”), and 3 titles appeared to be entirely unrelated to the others (“Radiopacity of Clomipramine Conglomerations and Unsuccessful Endoscopy: Report of 4 Cases,” “Intoxication with a Tricyclic Antidepressant,” and “Acute Poisoning by New Psychotropic Drugs”).

In this example, one can discern two overlapping and closely related main findings or “nuggets”: (a) zolpidem is associated with potential for abuse and dependence, and (b) acute withdrawal of zolpidem can lead to seizures. Each of these main findings was sufficiently newsworthy to be the subject of multiple case reports on its own. Also, the close relationship of the two was explicitly reported in the titles of several case reports (acute withdrawal occurs in the context of dependence), so that it was clear that the two nuggets really expressed different sides of a larger, single main finding or super-nugget. Also, note that in this example, tabulating the co-occurrence frequencies of title words was sufficient to discern the major findings (Table 2, online only). The four words (zolpidem, seizure, withdrawal, dependence) form a single clique in which all words co-occur pairwise in at least 4 titles. This suggests that there may be some type of close relation between cliques and nuggets. However, the relationship requires further study, since not all cliques represent main findings and not all nuggets are represented by cliques (see "Analysis of randomly chosen Medical Subject Headings–based PubMed queries" below).

Additional examples of multiple case reports were found that related drugs to disorders (e.g., topiramate as a rare cause of psychosis). We also found examples by looking for reports in which unexpected improvement occurred. Using the search query [(unexpectedly OR surprisingly OR unexpected OR surprising) AND (improved OR improvement OR ameliorated OR amelioration OR recovered OR recovery) AND (case reports[PT] OR (english abstract[PT] AND case[TIAB]))], we found eight reports of spontaneous or induced recovery after long-term vegetative state or coma. We also found a separate group of case reports in which soldiers were given prazosin for benign prostatic hypertrophy, which unexpectedly helped their symptoms of post-traumatic stress disorder. These open-ended searches demonstrated that nuggets of significant size did exist in the case report literature and that at least some could be readily found by examining patterns of title word usage across articles.

Analysis of randomly chosen Medical Subject Headings–based PubMed queries

To examine nuggets more systematically and facilitate manual inspection of each entire query set, we randomly selected PubMed queries defined using a MeSH term (restricted to case reports), each of which retrieved 15–35 articles. Table 3 (online only) shows the results for 30 query sets. Almost half of the literatures contained a nugget, and a sixth of them contained 8 or more articles that shared the same main finding. This is especially striking given that the literatures consisted of only 20.3 articles on average! The query [Plakophilins[MeSH] AND case reports[PT]] generated 2 distinct main findings: 1 concerned plakophilin-1 mutations that cause ectodermal dysplasia skin fragility syndrome, and 1 reported novel plakophilin-2 mutations. These were associated with distinct sets of title word co-occurrences [mutation, plakophilin 1/PKP1, gene, ectodermal, dysplasia, skin, fragility, syndrome] and [novel, mutation, plakophilin-2/PKP2].

A particularly striking example generating two distinct nuggets was ["shiitake mushrooms"[MeSH] AND case reports[PT]], which retrieved thirty-three articles on May 6, 2014. Of twenty-five articles that mentioned the word "shiitake" in the title, two subsets of articles reported different main findings: Four articles reported the occurrence of hypersensitivity pneumonitis caused by exposure to shiitake mushroom spores. Separately, seventeen articles reported the occurrence (or features) of dermatitis caused by eating shiitake mushrooms. Some of these discussed the distinctive flagellate erythema associated with the condition, whereas others emphasized clinical variability of the syndrome or flagged the novel geography of this clinical presentation. The main findings were immediately apparent from reading the titles: for example, "Clinical Variability of Shiitake Dermatitis"; "Shiitake Dermatitis: The First Case Reported from a European Country"; and "Shiitake Mushroom-Induced Flagellate Erythema: A Striking Case and Review of the Literature."

To make an initial assessment of whether cliques of title words generally point to the existence of nuggets within a query set, the number and features of cliques were calculated for each MeSH query set and compared to the manually identified nuggets. Neither the presence of cliques nor the number of cliques in a given query set correlated well with the presence or size of a nugget (Table 3, online only: $r = \sim 0.2$ for either clique presence or clique number vs. number of articles that map to the nugget). Thus, better term processing, more stringent definitions of cliques, and possibly other techniques entirely will be needed to identify nuggets in an automated fashion.

DISCUSSION

The clinical case report literature appears to be a promising test bed for attempting to identify "nuggets" (i.e., subsets of articles that state the same or closely related main finding). We confirmed that the main finding of a case report is often directly stated in its title. Nuggets are surprisingly prevalent and can be surprisingly large—the largest found so far is one finding reported in seventeen articles. Sometimes several overlapping related nuggets (as detected in titles) fit together to form a super-nugget that expresses a more complex main finding.

Focusing on shared title words will miss some nuggets. For example, in the commotio cordis query set (Table 3, online only), the occurrence of commotio cordis was reported in four articles related to violence, each of which employed different title terms (violence, less-lethal weapon, soldier, military). There is also some subjectivity in deciding whether two findings are essentially identical; for example, shiitake dermatitis was reported in three distinct titles as appearing in France, Spain, and "a European country." At one level of granularity, these are saying the same thing: a disease that was previously restricted to Asia was now being seen in Europe. Yet without knowing the full medical and geographical context of these reports, it is difficult to decide if these semantically similar title terms should be placed together. Another important limitation is the fact that the main finding of an article is sometimes stated only in its abstract or full text.

We plan to utilize the PubMed search query sets as gold standards for rule-based or machine learning approaches to predict which query sets contain nuggets and to identify and construct nuggets automatically. As a long-term goal, we aim to create a public web-based tool that, for a given PubMed case report query, identifies its nuggets, identifies case report articles that map to them (based on both title and abstracts), and summarizes the findings in graphical form. This would represent a prototype of finding-based information retrieval of the biomedical literature, extending current functionalities that are primarily based on finding articles devoted to a given topic.

Limitations

Necessary decisions were made in this study, which were often subjective and therefore subject to bias. Confirmation bias cannot be ruled out. This study was based on forty-five PubMed queries and might not represent the universe of PubMed-indexed case reports.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health grants R01 LM010817 and P01 AG039347.

REFERENCES

1. Aggarwal CC, Zhai CX. Mining text data. New York, NY: Springer; 2012.
2. Nieder C, Pawinski A, Dalhaug A. Contribution of case reports to glioblastoma research: systematic review and analysis of pattern of citation. Br J Neurosurg. 2012 Dec; 26(6):809–12.
3. Nabil S, Samman N. The impact of case reports in oral and maxillofacial surgery. Int J Oral Maxillofac Surg. 2012 Jul;41(7):789–96.
4. Centers for Disease Control and Prevention (CDC). Pneumocystis pneumonia—Los Angeles. MMWR Morb Mortal Wkly Rep. 1981 Jun 5;30(21):250–2.
5. Gottlieb GJ, Ragaz A, Vogel JV, Friedman-Kien A, Rywlin AM, Weiner EA, Ackerman AB. A preliminary communication on extensively disseminated Kaposi's sarcoma in young homosexual men. Am J Dermatopathol. 1981 Summer;3(2):111–4.
6. Hymes KB, Cheung T, Greene JB, Prose NS, Marcus A, Ballard H, William DC, Laubenstein LJ. Kaposi's sarcoma in homosexual men: a report of eight cases. Lancet. 1981 Sep 19;2(8247):598–600.
7. Barić H, Andrijašević L. Why should medical editors CARE about case reports? Croat Med J. 2013 Dec;54(6): 507–9.
8. Balistreri WF. Notice: an overlap in case reports. J Pediatr. 2008 Feb;152(2):295; author reply 295.
9. Rosenthal DI. What makes a case report publishable? Skeletal Radiol. 2006 Sep;35(9):627–8.
10. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. J Biomed Inform. 2012 Oct;45(5):885–92.
11. Smalheiser NR, Zhou W, Torvik VI. Anne O'Tate: a tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab. 2008 Feb 15;3:2.
12. The Arrowsmith Project. Anne O'Tate search engine [Internet]. University of Illinois at Chicago [cited 17 Feb 2015]. <http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi>.
13. Stanford Natural Language Processing Group. Stanford CoreNLP package version 3.5.1 [Internet]. The

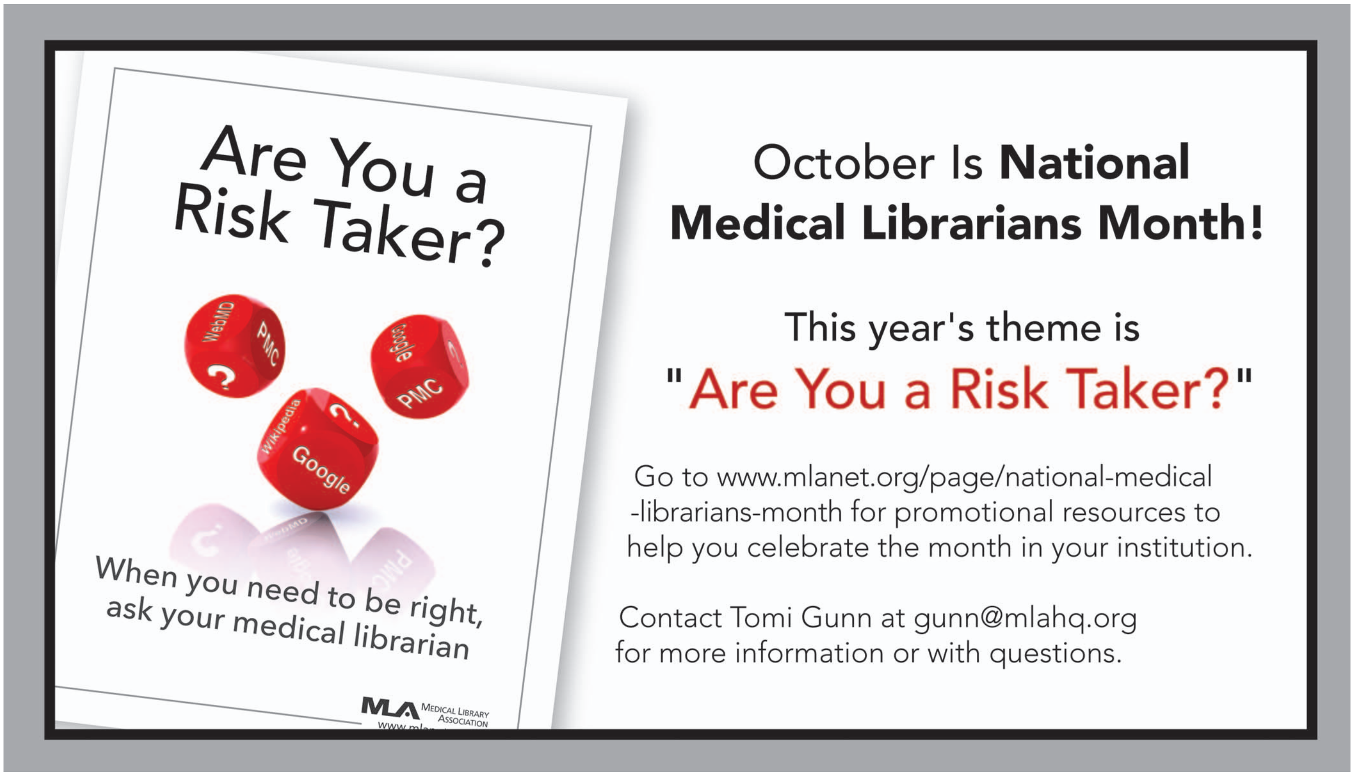
Group [cited 17 Feb 2015]. <http://nlp.stanford.edu/software/corenlp.shtml>.

14. Doyle D. English stopwords [Internet]. Ranks NL [cited 17 Feb 2015]. <http://www.ranks.nl/stopwords>.

AUTHORS' AFFILIATIONS

Neil R. Smalheiser, MD, PhD (corresponding author), neils@uic.edu, Associate Professor, Department of Psychiatry; **Weixiang Shao, MS,** software.shao@gmail.com, PhD Candidate, Department of Computer Science; **Philip S. Yu, PhD,** psyu@uic.edu, Distinguished Professor and Wexler Chair in Information Technology, Department of Computer Science; University of Illinois at Chicago, Chicago, IL 60612

Received February 2015; accepted May 2015


Are You a Risk Taker?
WebMD
PMC
?
Google
PMC
Wikipedia
?
Google
When you need to be right, ask your medical librarian
MLA Medical Library Association