Titles of Biomedical Articles: a corpus-based analysis

Brian Budgell
Canadian Memorial Chiropractic College
Toronto, Canada

Abstract

Background: Despite the importance of the title to the biomedical article, little quantitative information exists to guide authors and editors in formulating and evaluating titles. Furthermore, modern healthcare practices and policy making require the efficient recovery of clinically relevant information from the mass of electronic resources available. The retrieval of manuscripts for the purposes of informing policy and clinical decision making is strongly influenced by the manuscript title and abstract; hence, the importance of concentrating unambiguous clinically relevant information within these portions of an article. This imperative is especially pronounced in the instance of randomized controlled trials, the most rigorous form of original research into clinical effects of interventions. This article describes the application of the methods of corpus linguistics to the quantitative study of the titles of articles published in leading medical journals, with a special emphasis on titles of randomized controlled trials.

Methods: Titles were extracted from 2 corpora: the first consisting of 1,000 articles of various genre from the 4 leading medical journals, the second consisting of 310 randomized controlled trials (RCTs). Both corpora were analyzed for commonly occurring words, phrases and formats. The frequencies of the most commonly occurring words and formats in the 2 title corpora were compared. The titles of RCTs were also searched for clinically important information as determined by reference to the CONSORT checklist of content items for randomized controlled trials.

Results: Titles of biomedical articles are characterized by distinct conventions of word choice, length, recurrent phrases and format. All of these characteristics appear to be influenced by genre of article. For example, titles of RCTs contain approximately twice as many tokens as non-RCT titles. Additionally, format preference varies according to journal. In the leading medical journals, titles of RCTs use a constrained but highly technical vocabulary which makes frequent reference to research methodology, the treatment(s) under investigation, and the target disease.

Discussion: Quantitative analysis of 2 corpora of biomedical titles has revealed distinctive lexical and syntactical features which could aid in human and machine-based knowledge extraction. Titles of RCTs, in particular, are a dense source of unambiguous clinically important information and so are illuminating in and of themselves. Additionally, titles may act as a Rosetta Stone in deciphering abstracts and bodies of full articles.

Introduction

Titles of medical journal articles fulfill a number of important purposes including assisting potential readers in identifying and retrieving papers most relevant to their interests (1). Thus, the International Committee of Medical Journal Editors (www.icmje.org) recommends that a title should accurately and concisely reflect the contents of the article. In order to improve the comprehensibility of medical articles, a number of more genre-specific content guidelines have been published and subsequently endorsed by many medical journals. A comprehensive and up-to-date listing of medical content guidelines is maintained by the Equator Network (www.equator-network.org). Among the earliest and most important content guidelines, CONSORT - the Consolidated Standards of Reporting Trials - makes reference to the wording of titles (2), as does STROBE – Strengthening the Reporting of Observational Studies in Epidemiology (3). Nonetheless, the effects of these various guidelines on the formulations of titles have not been documented.

Content guidelines are designed to enhance the comprehensibility of biomedical manuscripts for human readers. Within the context of health care, this is an urgent concern, since clinical reports are published with the primary intention of improving health care outcomes. This is achieved by influencing the clinical decision making of individual reader-clinicians, and by influencing health care policy. This latter function is increasingly affected by the meta-analysis of multiple clinical trials, as for example via the Cochrane Collaboration (www.cochrane.org), and the subsequent formulation of policies, guidelines and standards. Given the rapid growth of clinical research, meta-analysis of primary data sources cannot keep pace with knowledge growth unless computer-assisted methodologies are employed.

To date, computer-based systems for extraction of information from biomedical manuscripts have focused on molecular biology (4-7). Notwithstanding the rigour of this work, the outcomes are, by and large, distant from the clinical interface. Hence, there is a need for computer-based methodologies which assist in making clinically important knowledge readily comprehensible to end-users.

Recently the methods of corpus linguistics have begun to be applied to analysis of the clinical literature (see for example 8-12), providing quantitative information about biomedical language within certain health disciplines. Particularly as medical publication occurs in what for most authors is a second language, a more explicit understanding of the extant standards for titles, as derived by objective methods, would almost certainly facilitate the processes of writing, editing and information extraction. Thus, this study was undertaken to gather baseline data on the lexical and syntactical features of titles of biomedical articles. Armed with this information, authors and others may make evidence-based decisions concerning the "attestedness" of titles which they create or encounter. Furthermore, objective data on current language practices should illuminate efforts to develop computer-based systems to extract knowledge from biomedical texts.

Thus the present study analyzed two biomedical corpora: a non-specific corpus consisting of titles of 1,000 items of various genres, and a genre-specific (RCT) corpus of titles of 310 randomized controlled trials. Titles were analyzed for tokens, types, format, and n-grams (recurrent phrases).

Methods

A search of PubMed was performed on January 27, 2008 for the 250 most recently indexed entries from each of 4 journals: The British Medical Journal (BMJ), The Journal of the American Medical Association (JAMA), The Lancet (LANCET), and The New England Journal of Medicine (NEJM). These journals were chosen because they are widely read and cited (most recent impact factors 9.723, 25.5, 28.6 and 51.296, respectively), they publish on a broad range of topics, and they include a variety of genre, such as letters, commentaries, news items and clinical research reports. NEJM and JAMA follow the conventions of American English, whereas BMJ and Lancet use British English. NEJM has been the subject of a smaller corpus-based study of titles (11). This non-specific corpus did not distinguish between genre of publication (e.g. interventional versus observational study), but does permit comparisons between journals.

The genre-specific corpus of titles of randomized controlled clinical trials was created on October 14, 2008 by searching PubMed with the search string *"BMJ (Clinical research ed.)"[Jour] OR "Lancet"[Jour] OR "JAMA : the journal of the American Medical Association"[Jour] OR "The New England journal of medicine"[Jour] OR "Annals of internal medicine"[Jour] Limits: Publication Date from 2005/01/01 to 2005/12/31, Randomized Controlled Trial, English.* Hence, the search did not specify the use of the term "random" or any derivative thereof in the manuscript title. The corpus of RCTs was created as part of a separate study of biomedical language and so included Annals of Internal Medicine (impact factor 15.5), which was not included in the non-specific corpus of 1,000 titles. The search retrieved all RCTs published in the target journals in the calendar year 2005. This date restriction was used to allow comparisons (not reported herein) to corpora in nursing (8), public health (9) and midwifery and perinatal care (12) which were also limited to articles published in the year 2005.

The corpora were analyzed with the software WordSmith Tools 5.0 (Oxford University Press, Oxford). For each title in the two corpora, the number of tokens was recorded. In this study, hyphenated phrases were considered single tokens if the components lost meaning when separated; for example *non-invasive* would be considered a single token, whereas *34-year-old* would be considered 3 tokens. The types (different words) in the two corpora were classified as being from i) the General Service List (GSL) – the 2,000 most common word families in written English (13), ii) the Academic Word List (AWL) – the additional 570 word families commonly encountered in academic environments (14), or being off-list i.e. on neither of the 2 preceding lists. It has been argued that off-list types are likely to be particular to or have specific nuances within a target corpus (15).

Duplicate titles were eliminated from the non-specific corpus of 1000 titles (duplicates normally signal correspondence concerning a single article) and then the remaining 913 titles of the non-specific corpus and the 310 titles of the genre-specific corpus (no duplicates) were analyzed to identify n-grams, i.e. commonly recurring phrases. The formats of titles were classified as single phrases, phrase-compounds, statements, statement-compounds, questions or question-compounds. A single phrase was defined as a single word or a string of words which expressed a research theme but which did not constitute a complete sentence; e.g. *Caring for people with*

dementia. A phrase-compound was defined as a series of (usually 2) phrases; e.g. *Pilocarpine: better than a scan*. A statement was defined as a declaration of fact consisting of a subject, verb and (if appropriate) direct object, indirect object or complement; e.g. *Prime minister promises raft of new screening tests.* A statement-compound consisted of a statement followed by one or more additional phrases or sentences; e.g. *Researchers deconstruct metastasis: genetic clues revealed*. A question was a complete interogative sentence; e.g. *Does this child have a urinary tract infection*? A question-compound was a question followed by one or more additional phrases or sentences; e.g. *What works? Interventions for maternal and child undernutrition and survival.*

Results

For the non-specific corpus, 1,000 titles were recovered consisting of 8410 tokens and 2565 types. The mean title lengths (and 95% confidence intervals) for the four journals were BMJ: 8.16 (7.59-8.73), JAMA: 9.16 (8.70-9.62), Lancet: 8.00 (7.32-8.68), and NEJM: 8.14 (7.64-8.64). The genre-specific (RCT) corpus consisted of 5,183 tokens made up of 1,368 types. The mean length (and 95% confidence interval) of the 310 titles was 15.6 (14.95-16.23) tokens. Hence, the mean length of titles of RCTs was significantly longer than the mean length of titles in the non-specific corpus ($z = -19.109$, $p < 0.001$ per Wilcoxon rank sum test). The numbers of tokens per title in the RCT genre-specific corpus and in the non-specific corpus are presented in figure 1.

For the total 1,000 titles in the non-specific corpus, 59% of the tokens were from the GSL, 9% were from the AWL, and 32% were off-list (neither GSL nor AWL). Types occurring with a frequency of greater than 1/1,000 tokens are listed in table 1. Types occurring at a frequency of greater than 1/100 tokens included 9 types from the GSL which collectively made up approximately 22% of the corpus. These 9 highly prevalent words were *of, and, the, for, a, to, with* and *on*. The 10 most common "content words" were *health, patients, clinical, risk, cancer, care, trial, disease, study* and *treatment*. The pronoun, *I*, was absent and *we* occurred only 7 times.

For the genre-specific corpus of 310 RCTs, 54% of tokens were either numbers or words from the GSL, 5.8% of tokens were from the AWL, and 39.5% of tokens were off-list words. The ten most common content words (and their raw frequencies) were *trial* (x210), *controlled* (x130), *randomised* (x122), *randomized* (x76), *patients* (x58), *treatment* (x47), *effect* (x35), *disease* (x30), *therapy* (x29), and *study* (x25). The ten most common 3- or 4-word recurring phrases (n-grams) in the non-specific corpus are listed in table 2. The two most common meaningful 3-grams (and their raw frequencies) in the genre-specific (RCT) corpus were *randomised controlled trial* (x75) and *randomized controlled trial* (x7). The two most common 2-grams (and their raw frequencies) were *randomized trial* (x34) and *randomised trial* (x24).

The distributions of formats in the complete 1,000 title non-specific corpus and the genre-specific (RCT) are presented as percentages in figure 2. Overall, the distributions of formats in the non-specific corpus were phrase: 54%, phrase-compound: 29%, statement: 14%, statement-compound: 0%, question: 2% and question-compound: 1%. For the RCT corpus, the

corresponding distributions were phrase: 38%, phrase-compound: 61%, statement: 0%, statement-compound: 0%, question: 0% and question-compound: 1%. The differences between distributions of formats in the two corpora are statistically significant ($X^2 = 127.7041$, df = 4, $p < 0.001$ per Chi squared test).

With regard to clinically relevant information, of the 310 titles of RCTs, 280 (90%) made reference to a health condition, e.g. *Crohn's disease, atrial fibrillation*. Furthermore, 279 titles (90%) named at least one intervention, e.g. *radiotherapy, antibacterial prophylaxis* and 107 titles (35%) made reference to more than one intervention. However, of these, only 73 titles (24%) explicitly indicated a head-to-head comparison of 2 or more interventions. Only 31 titles (10%) explicitly identified an outcome measure, e.g. *3-year disease-free survival*.

Discussion

This study presents a quantitative analysis of 2 biomedical corpora: a non-specific corpus of 1,000 titles of articles from a range of genre published in four leading biomedical journals, and a genre-specific (RCT) corpus consisting of titles of 310 randomized controlled trials from recent issues (2005) of the 5 leading biomedical journals. Previous linguistic studies have generally not distinguished between interventional medical studies, such as randomized controlled trials, and observational studies, despite the fact that these different genres take quite different approaches towards informing health care practices.

The average length of titles in the non-specific corpus ranged from 8.0 tokens for Lancet to 9.16 tokens for JAMA. These small differences may be accounted for by the different distributions of genres in the four journals, something which this study did not examine. Furthermore, despite the strong central tendency, the number of tokens per title actually ranged from 1 to 44. Titles of RCTs were substantially longer, averaging 15.6 tokens and ranging from 5 to 37 tokens. Fifty-nine percent of the tokens in the non-specific corpus were from the GSL, 9% were from the AWL, and 32% were not on either of these lists. In the genre-specific corpus, 54% of tokens were from the GSL, 5.8% were from the AWL, and 39.5% were off-list. The GSL normally represents approximately 80% of general English texts (16). The AWL was previously reported to represent approximately 10% of tokens in texts of medical research articles (17). The relatively high proportion of off-list tokens in the two corpora examined herein demonstrates that in general, and specifically for RCTs, titles of medical articles are likely to be particularly inaccessible to persons without specialist knowledge of medical vocabulary (15).

As in general English, a very few function words made up more than 20% of the non-specific corpus. The most common type in the GSL, *the*, represented less than 3% of tokens from the non-specific corpus, versus 7% in general English texts (16). This is consistent with the obvious preference for brevity over grammatical correctness in title formulation; that is to say that function words such as *a*, *an* and *the* were frequently omitted from titles. The absence of the common pronoun *I* and the under-representation of *we* are consistent with the virtual absence of self-reference in the texts of biomedical articles, and this appears to be a convention of scientific writing (18, 19). The token/type ratio was 3.25 for the non-specific corpus versus 3.79 for the

genre-specific corpus reflecting the recurrent usage of a few terms, such as *trial*, *controlled* and *randomized*.

The recurrence of certain fixed phrases and the relatively high frequency of their reference to experimental design is consistent with the high prevalence of experimental reports and reviews in biomedical journals. Two hundred and nine of the 310 titles in the genre-specific (RCT) corpus (67%) from the year 2005 made reference to study design, as recommended in the most recent extension of the CONSORT statement, the most broadly accepted content guidelines for reports of randomized controlled trials (2). In this regard and among the 10 most frequently occurring words in the RCT corpus, the terms *randomized*, *randomised*, *trial* and *study* were completely unambiguous, being used only to describe experimental design. Similarly, in 129 of 130 occurrences, the term *controlled* was used to refer to experimental design, meaning that the design included a group of patients who received a standard treatment to which the effects of an experimental treatment could be compared. The exceptional instance of usage was "…*patients with suboptimally controlled type 2 diabetes*…". The term *patients* was always used in the sense of a person receiving or requiring the medical intervention described in the paper. *Treatment* and *therapy* were always used in the sense of a medical intervention to promote health or combat disease. The term *effect* was only used as a noun and in the sense of an outcome of a disease or intervention. The term *disease* was only used in the sense of the specific pathological entity for which treatment was described in the paper, and was most often preceded by a descriptor, e.g. *cardiovascular disease, Crohn`s disease*. Furthermore, when an intervention (treatment or therapy) and a disease appeared together in a title, they bore an unambiguous relationship – the intervention was directed against that specific disease. Similarly, if a cohort of participants and a disease appeared within the same title, the patients were afflicted with that specific disease. Thus there was remarkably little ambiguity in the most commonly recurring terms; terms which were used to signal clinically important information as defined by the CONSORT statement (2).

In an analysis of a smaller corpus, titles of biomedical articles were previously characterized as nominal (equivalent to "phrase" in this study), compound ("phrase-compound"), full sentence ("statement") and question (10). In the present study, the range of classifications was increased to allow for compound titles in which the first components were statements or questions. It was previously reported that the nominal (phrase) and compound (phrase-compound) formats predominate in the field of medicine (10), and the present findings bear this out, with approximately 83% of titles in the non-specific corpus using one of these forms. It had previously been reported that NEJM displayed a strong preference for the phrase (nominal) style, and the present study substantiates this. BMJ, on the other hand, seemed tolerant of diversity in formats of titles.

In the genre-specific (RCT) corpus, 118 titles (38%) used the phrase (nominal) format and 190 titles (61%) used the phrase-compound format. The phrase format often followed the pattern ***treatment A*** *for* ***disease/condition B***. Hence, a representative RCT title in the phrase format would be "*Infliximab for induction and maintenance therapy for ulcerative colitis.*" The first phrase of the phrase-compound title is often used to declare the theme of the article, while the second phrase acts as a delimiter or modifier. In the non-specific corpus, approximately 4% of the titles employed a phrase-compound format in which the second phrase specified the experimental design; hence the high frequency of phrases such as *randomised controlled trial*

and *a systematic review*. In the genre-specific corpus, all but 1 of 190 phrase-compound titles specified the experimental design. Hence, a representative RCT title would be "*Plasma exchange when myeloma presents as acute renal failure: a randomized, controlled trial.*" As readers with an interest in biomedical articles frequently search by experimental design, this style of title would likely contribute significantly to the efficiency of article recovery.

The use of statements and questions as titles is rare in other domains (10), and still apparently uncommon for biomedical articles. While the length may be problematic, a full sentence or question is likely to be less ambiguous than a phrase or phrase-compound title. On the other hand, a statement, which is necessarily brief in a title, may leave an author vulnerable to criticism that the title does not accurately represent the findings of the study – for example, a simple statement that the experimental treatment was indeed effective against the target disease might miss all sorts of important caveats such as the demographics of the patients. Among the 310 RCT titles, only 2 used a question-compound format, and no titles were in the form of a statement or statement-compound. Those few statements and questions which did appear in the two corpora were not constrained by the normal conventions of grammar, although there were no apparent errors in spelling nor in agreement between subjects and their verbs. In this regard, the BMJ and Lancet appeared to make occasional attempts at word play; e.g. "*Dissent of the testis*." JAMA and NEJM titles were devoid of humour, and there were no apparent attempts of humour in the genre-specific (RCT) corpus.

In summary, this analysis demonstrates that titles from leading medical journals use a restricted number of formats and a constrained vocabulary which is likely to be indecipherable to those without specialist knowledge of the field. Overall, some journals appear to have preferences for particular formats. Furthermore, there are apparently genre-specific biases in format and phraseology. Notwithstanding this overall uniformity, exceptional formats were still published, suggesting that authors enjoy some freedom to be inventive.

With regard to knowledge extraction, titles of RCTs represent a relatively rich source of remarkably unambiguous, clinically important information, and so may act as something of a Rosetta Stone for more ambiguous portions of a manuscript. Specifically, more than two thirds of titles explicitly stated design (CONSORT checklist item #1), 90% unambiguously indicated the disease or health condition of interest (CONSORT checklist item #3a), and 90% specified one or more interventions (CONSORT checklist item #4). As a recent extension of the CONSORT Statement has more explicitly stipulated standards for titles of reports of RCTs (2), and as the 5 leading biomedical journals used for the RCT corpus (and numerous other biomedical journals) have endorsed this statement as policy (20), it is reasonable to expect that titles of RCTs will become an even more reliable source of clinically important information.

Acknowledgement: This work was supported by grant # 0058/07 from GDS International. David Soave provided advice on statistical analysis.

References

1. Demner-Fushman D, Hauser S, Thoma G. The role of title, metadata and abstract in identifying clinically relevant journal articles. In: AMIA 2005; 2005; Washington, D.C.: America Medical Informatics Association; 2005. p. 191-195.
2. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, Schulz KF; CONSORT Group. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. PLoS Med 2008 Jan 22;5(1):e20.
3. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Ann Intern Med 2007; 147(8):573-577.
4. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus – a semantically annotated corpus for bio-textmining. Bioinformatics 2003;19(Suppl 1):i180-i182.
5. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics 2005;6(Suppl 1):53.
6. Cohen KB, Fox L, Ogren PV, Hunter L. Empirical data on corpus design and usage in biomedical natural language processing. AMIA Annual Symposium Proceedings 2005:156-160.
7. Pyysalo S, Ginter F, Heimonen J, Bjorne J, Boberg J, Jarvinen J, Salaksoki T. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics 2007;8:50.
8. Budgell B, Miyazaki M, O'Brien M, Perkins R, Tanaka Y. Developing a corpus of the nursing literature: a pilot study. Japan Journal of Nursing Science 2007;4:21-25.
9. Millar N, Budgell B. The language of public health – a corpus based analysis. The Journal of Public Health 2008;16(5):369-374.
10. Soler V. Writing titles in science: an exploratory study. English for Specific Purposes 2007;26:90-102.
11. Wang Y, Bai Y. A corpus-based syntactic study of medical research article titles. System 2007;34:388-399.
12. Chiba Y, Millar N, Budgell B. The language of midwifery and perinatal care: a quantitative analysis. Journal of Japan Academy of Midwifery 2010;24 (1) 74-83.
13. West MP. A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology. Longmans Green, London; 1953.
14. Coxhead A. A new academic word list. TESOL Q 2000;34:213–238.
15. Chung T, Nation P. Identifying technical vocabulary. System 2004;32:251-263.
16. Nation P, Waring R. Word frequency and vocabulary size. In: Schmidt N, McCarthy M, editors. Vocabulary: Description, Acquisition and Pedagogy. Cambridge: Cambridge University Press; 1997. p. 6-19.
17. Chen Q, Ge G. A corpus-based lexical study on frequency and distribution of Coxhead`s AWL word families in medical research articles (RAs). English for Specific Purposes 2007.
18. Hyland K. Humble servants of the discipline? Self-mention in research articles. English for Specific Purposes 2001;20:207-226.
19. Martinez I. Impersonality in the research article as revealed by analysis of the transitivity structure. English for Specific Purposes 2001;20:227-247.
20. Hopewell S, Altman DG, Moher D, Schulz KF. Endorsement of the CONSORT Statement by high impact factor medical journals: a survey of journal editors and journal 'Instructions to Authors'. Trials 2008;9:20

Figure 1. Distributions of title lengths (number of tokens per title) expressed as a percentage of the total number of titles in each corpus and from each individual journal within the non-specific corpus.

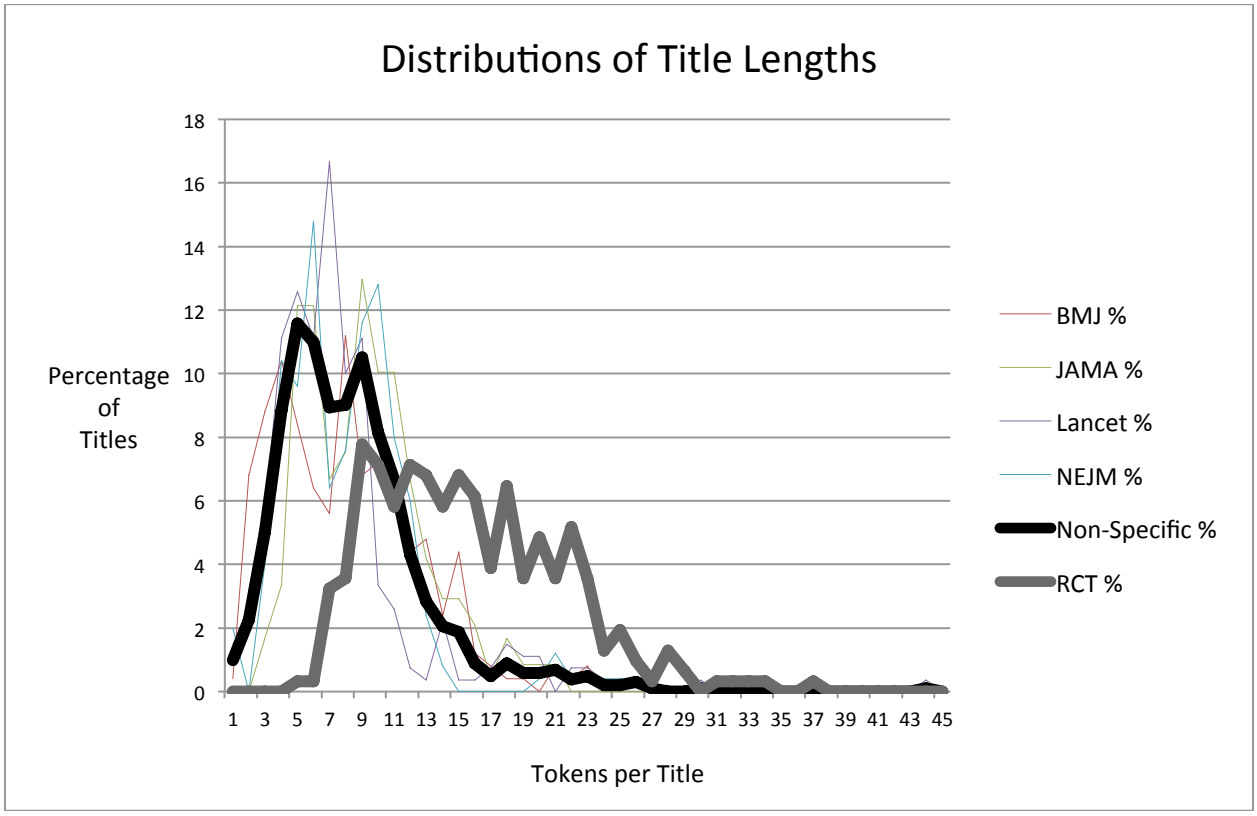

Figure 2. Distributions of title formats in non-specific vs RCT corpus

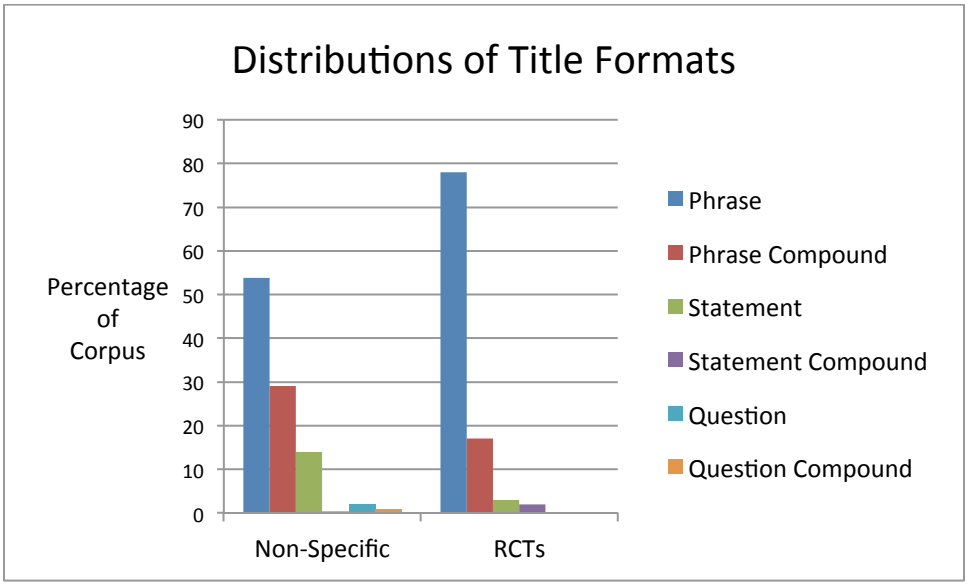

Table 1. Types with frequency greater than 1/1,000 tokens in non-specific corpus

Type	Frequency	List	Type	Frequency	List
of	361	GSL	heart	15	GSL
and	276	GSL	high	15	GSL
the	251	GSL	it	15	GSL
for	184	GSL	safety	15	GSL
a	149	GSL	use	15	GSL
to	129	GSL	cardiovascular	14	off-list
with	95	GSL	drugs	14	off-list
health	77	GSL	effects	14	GSL
on	53	GSL	low	13	GSL
patients	53	off-list	page	13	GSL
clinical	44	off-list	response	13	AWL
risk	44	GSL	severe	13	GSL
cancer	43	off-list	associated	12	GSL
care	40	GSL	cover	12	GSL
trial	40	GSL	man	12	GSL
disease	36	GSL	time	12	GSL
study	34	GSL	human	11	GSL
treatment	32	GSL	people	11	GSL
medicine	28	GSL	randomized	11	off-list
chronic	26	off-list	report	11	GSL
controlled	25	GSL	review	11	GSL
is	25	GSL	trials	11	GSL
randomised	25	off-list	brain	10	GSL
year	24	GSL	effect	10	GSL
from	23	GSL	interventions	10	AWL
new	23	GSL	management	10	GSL
patient	23	GSL	maternal	10	off-list
case	22	GSL	meta	10	off-list
coronary	22	off-list	outcomes	10	AWL
mortality	22	off-list	quality	10	GSL
breast	21	off-list	renal	10	off-list
global	21	AWL	UK	10	off-list
medical	20	AWL	what	10	GSL
acute	19	off-list	arthritis	9	off-list
children	19	GSL	call	9	GSL
or	19	GSL	doctors	9	GSL
failure	18	GSL	general	9	GSL
diabetes	17	off-list	injury	9	AWL
hospital	17	GSL	mind	9	GSL
images	17	AWL	my	9	GSL
old	17	GSL	piece	9	GSL
analysis	16	AWL	plus	9	AWL
at	16	GSL	prostate	9	off-list
be	16	GSL	records	9	GSL
older	16	GSL	reduce	9	GSL
primary	16	AWL	should	9	GSL
screening	16	GSL	syndrome	9	off-list
US	16	GSL	systematic	9	GSL
versus	16	off-list	treating	9	GSL
by	15	GSL	war	9	GSL
genetic	15	off-list	women	9	GSL

Table 2. Ten most common 3- and 4-grams in non-specific corpus

Phrase	**Frequency**
in patients with	18
randomised controlled trial	11
randomized controlled trial	11
in the United	8
a randomized controlled trial	7
for treatment of	6
in patients with HIV	5
a systematic review	5
of cardiovascular disease	5
and the risk of	4