Original articles

The increasing pseudodignification of medical prose

Neville W Goodman
Retired Consultant Anaesthetist, Bristol, UK; nevwgoodman@mac.com

Abstract

Databases such as PubMed and programmes such as Google's Ngram Viewer allow the measurement of the prevalence of words and how those prevalences have changed over time. They also allow comparisons of the uses of words in medical English and more general English.

Many words that are disapproved of in style guides increased in prevalence between 1975 and 2010 ("up-words"). Some, such as *option* or *options* and *impacts* (as a verb), were up-words also on Ngram; but others, such as words with the root *address* (which have increased 27-fold in PubMed), were not up-words on Ngram.

In general, the ratio of the disapproved to the approved of any pair of synonyms was higher in PubMed than on Ngram: for example, the ratio of *raised* to *elevated* was 0.25 in PubMed and 20 on Ngram; the ratio of *given* to *administered* was 1.5 in PubMed and 57 on Ngram. The relative synonym use can be expressed as a single figure, that is, *raised* to *elevated* was 83 times greater on Ngram than in PubMed; and *given* to *administered* was 38 times greater. For most of the pairs of synonyms for which comparisons were possible, the difference between general English and medical English has increased since 1930, when *raised* to *elevated* was three times greater on Ngram, and *given* to *administered* nine times greater.

As long as medical authors gain nothing by writing more simply and clearly, it will be difficult to stop or reverse the increasing complexity of medical prose.

Keywords

Writing, linguistics, prevalence

Introduction

Using long words for short is a feature of what Michael O'Donnell termed the "pseudodignification" of medical prose[1]. Databases such as PubMed[2] and programmes such as Google's Ngram Viewer[3] allow the measurement of the prevalence of words and how those prevalences have changed over time. They also allow comparisons of the uses of words in medical English and more general English. When revising our style guide for medical English (*Medical Writing: a prescription for clarity*[4] – referred to here as "Prescription"), we knew which were the commonly used words in medical English, and we were able to compare – quantitatively – the way that usage has changed in medical English and in more general English. This paper is an extension of those comparisons and some of the discussion based on them.

PubMed is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. Except for some common words such as *showed* and *found*, which are known as "stopwords", any word can be searched and a count recorded for any year of publication. As an example, the searchable word *quantified* (an imprecise word, better replaced by *measured* or *counted*) occurred in 100 PubMed articles (strictly in the titles or abstracts of 100 articles published in English and whose abstracts are available to the database) in 1975, and in 5927 articles in 2010.

Ngram is an online programme that displays a graph showing how words (or phrases) have occurred in a corpus of books over selected years. The corpus we used in Prescription was "English", which includes technical works. For this article, I used "English fiction", to bring out better the differences between medical and more general English.

The aim of the study was to identify words whose prevalence altered between 1975, when abstracts first became readily available in PubMed, and 2010; in particular, to compare the use of shorter and longer similar words.

Methods

I searched PubMed for 2010, 1975, and 1921-1940 (see below); I used Ngram for 2008, 1975, and 1930. I used Microsoft Excel for all calculations.

The prevalences of the more common words were recorded as the number of PubMed titles or abstracts in which the words occurred (corrected for the total number of available articles in English and with available abstracts). Some words popular with medical authors have an obvious replacement, but the replacement is a stopword. To measure the relative popularity of stopwords, I compiled a file of 2000 abstracts from PubMed. I downloaded ten non-consecutive pages of 200 abstracts and deleted all the titles, author lists, and sources, so that the file was just of abstracts. Using Microsoft Word "find and replace" gave a count of the searched words, which I converted to a percentage (the file contains about 370,000 words). I also used this file to check the validity of using a count of abstracts containing words as an index of the prevalence of those words.

The precise form of a searched word varied with the word, and with the ease of search. PubMed accepts wildcards: thus, searching *quantif** will find all forms of the lexeme *quantify* (eg, *quantified*, *quantification*), but Ngram does not accept wildcards, so each form has to be searched separately. For many words, one or two forms predominate: eg, *quantif** retrieved ~18,000 medical articles from 2010, and each of *quantification* and *quantified* retrieved about ~6000. In the Discussion section, when a specific word is referred to by one of its forms, the precise search is indicated by the word's entry in a Table: eg, unless I refer specifically to *targeting*, *target* implies *target**.

PubMed, unlike Ngram, does not differentiate between parts of speech: thus, for example, a search cannot differentiate *impacts* as a noun and *impacts* as a verb (although scanning of the retrieved abstracts showed that most occurrences were as the verb).

For the synonyms (referred to as synonyms, although not strictly so) I attempted some limited comparison with language of an earlier era. I tried using PubMed articles published between 1901 and 1920, but there were only ~1450 articles with abstracts. For 1921-1940, there were 4202 with abstracts (out of a total of ~94,000), which I used to compare with Ngram for English fiction published in 1930.

In this article, I refer to any word whose prevalence increased as an "up-word", and to any whose prevalence decreased as a "down-word".

Results

Figure 1 shows, for the more common words, concordance between the number of abstracts that contain a word, and that word's count in the file of 2000 abstracts. (The outlier in Figure 1 – the point furthest right – is for the root *significan*: words with this root are likely to occur many times in any single abstract, so its position in the graph is not surprising.)

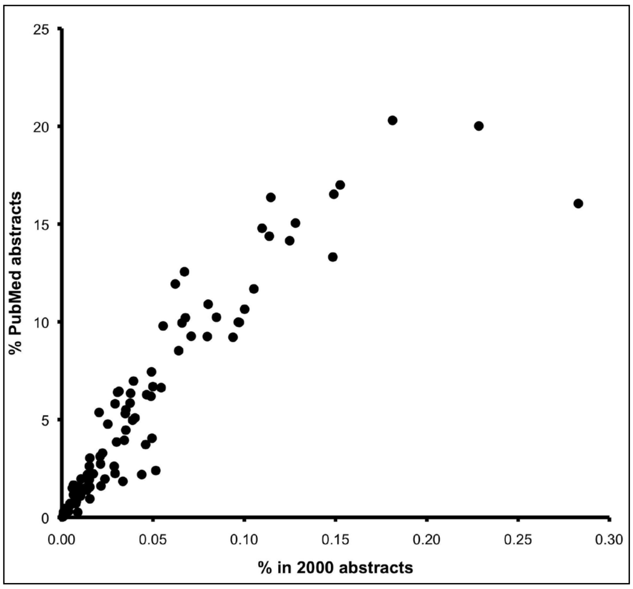

Fig 1. PubMed abstracts containing a word and prevalence of word in file of 2000 abstracts

Figure 2 shows the concordance between PubMed and Ngram of the commonest words in written English – this varies with source, but a reasonable first nine are *the*, the lexeme *be* (which includes *is*, *was*, etc), *to*, *of*, *and*, *a*, *in*, *that*, and the lexeme *have* (which includes *has* and *had*).

Table 1: Increases in prevalence of words in PubMed and on Ngram

Rank	Word	PubMed	Ngram
1	gender	43.6	5.5
2	address*	26.9	1.0
3	option/s	25.8	3.0
4	impacts	24.5	8.7
5	strategy/gies	21.9	1.4
6	novel	18.1	0.6
7	optimiz/sation	12.8	3.0
8	impact	11.9	1.0
9	mitig*	10.1	2.3
10	randomiz/sed	9.8	6.6
11	plethora	8.2	1.3
12	target*	7.6	1.5
13	quantification	7.2	1.3
14	cocktail	6.5	1.2
15	innovat*	6.2	0.9
16	paradigm*	5.8	1.4
17	tool/s	5.8	1.1
18	executed	5.5	0.9
19	myriad	5.4	1.1
20	quality	5.1	1.8
21	modulated	4.9	0.8
22	implicated	4.8	1.1
23	dearth	4.7	0.8
24	**worse**	4.3	1.2
25	controv*	4.2	0.8
26	amelior*	4.2	0.9
27	explor*	4.2	0.9
28	**allocated**	3.8	0.9
29	formally	3.8	0.8
30	recent/ly	3.7	0.9
31	feasible	3.6	0.6
32	modalit*	3.6	1.1
33	elucidat*	3.3	0.8
34	paramount	3.3	0.9
35	**affects**	3.1	0.9
36	attenuat*	3.1	0.7
37	dramatically	2.9	1.1
38	advocate/s	2.8	1.0
39	pathway*	2.8	1.5
40	provide/s	2.8	0.8

The 40 greatest increases (1975-2010) in prevalence of words looked at in PubMed by percentage of abstracts (in English, available abstract). Also shown are changes in prevalence on Ngram 1975-2008. Unity (1) is no change. Indexed entries NOT in Prescription are in bold. A slash indicates more than one word, eg, *option/s* is *option* and *options*. The PubMed wildcard * indicates all words starting with the root.

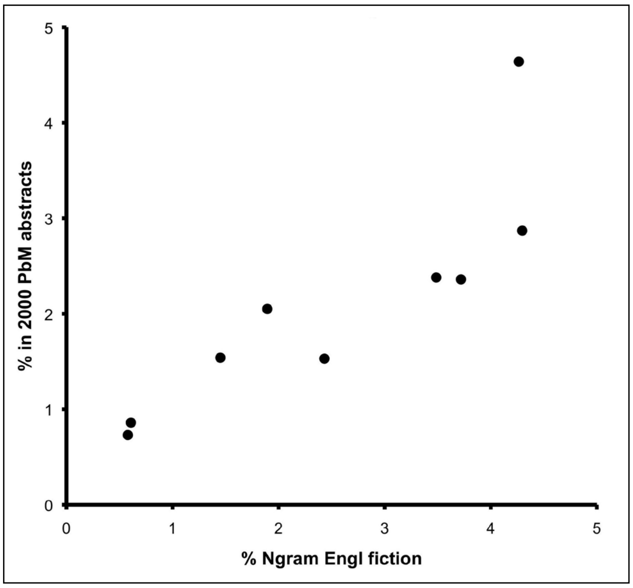

Fig 2. Prevalence of the commonest English words in PubMed and on Ngram

The 40 greatest up-words are shown in Table 1. Most of these words have a specific entry in Prescription. The position of a word in the Table depends on the precise form. The prevalence of the root *address* has increased 27-fold, but of the word *address* (which could be the infinitive, or the present indicative) has increased 44-fold. The root *target* has increased over 7-fold, but the participles *targeting* and *targeted*, together, have increased 177-fold.

The only words in Table 1 for which changes in prevalence on Ngram were more than 2-fold are all in the upper half: *gender*, *option* or *options*, *impacts* (verb) and *optimiz/sation* are all up-words. The prevalence of *novel* (adjective) decreased.

The greatest down-words on PubMed (not tabulated) were *basically*, *proved*, *former*, *noted*, *dosage*, *concerned*, *attained*, *latter*, and *instance* or *instances* (to less than 0.5-fold); and *probable* or *probably*, *essentially*, and *unimportant* (to less than 0.3-fold). All except *probable* or *probably* and *unimportant* have entries in Prescription.

I compared synonyms by comparing the ratios of use in PubMed and on Ngram. The ratio of *heart* to *cardiac* in PubMed is 1.2; on Ngram it is 365. Thus, the ratio comparison of the use of the two words shows that *heart* is 306 times more likely to be written than *cardiac* in more general English than in medical English. The closer a ratio comparison is to unity (1), the more similar are medical and general English: the ratio comparison for *above* and *below* is 1.33.

Table 2 is ranked by the most recent ratio comparisons: the upper table was calculated from the number of PubMed abstracts; in the lower table, the first synonyms are stopwords, and so the second column (1975) and third column (c.1930) are not available.

In general, the ratio comparisons are closer to unity in 1975, and closer again in 1930: five of the "disapproved" words did not even appear in any abstract in PubMed 1921-1940.

Table 2: Synonym ratios ranked by the comparison between Ngram 2008 and PubMed 2010

First synonym "approved"	Second synonym "disapproved"	Ngram: PubMed ratio comparison		
		2010	1975	c.1930
best	optimal/um	203.3	319.2	3948.9
raised	elevated	82.8	40.1	3.2
plans	strategies	75.8	7.7	2.8
new	novel	58.9	3.4	0.3
reached	attained/ achieved	50.0	11.8	5.1
given	administered	38.4	25.1	8.6
altered/ changed	modulated	36.3	5.1	-
explained	elucidated	32.9	29.8	5.7
give/s	provide/afford/ allow/s	32.6	33.1	2.1
idea/s	concept/s	31.3	27.5	64.8
few/lack	paucity/dearth	21.1	7.2	1.4
large/st	maximum/al	17.1	22.3	32.4
small/est	minimum/al	13.4	10.5	23.2
measured	evaluated	13.1	3.6	1.7
more	additional	11.3	6.9	5.7
affects	impacts_VERB	10.4	12.6	-
choice/s	option/s	7.6	1.1	-
sex	gender	6.6	0.7	-
proved	proven	5.9	2.8	3.1
suffered	experienced	5.6	2.4	0.9
measurement/s	quantification	3.4	0.7	-
many/numer-ous	plethora/myriad	2.1	0.6	2.8
suggested	advocated	1.9	1.1	2.0
effect	impact_NOUN	1.7	0.1	0.3
reduced	attenuated	1.1	0.4	1.4
did/done	performed	186.1	-	-
begin/start	commenced/ initiated	15.6	-	-
made	constructed	9.5	-	-
did/done	executed	7.7	-	-
showed	demonstrated	4.5	-	-
has/have	possess/es	4.0	-	-
most	majority	3.9	-	-
showed	exhibited	2.2	-	-
used	employed/ utiliz/sed	1.8	-	-
showed	revealed	1.5	-	-

Upper: using number of PubMed abstracts; - indicates second synonym did not occur in PubMed c.1930. Lower: using occurrence in file of 2000 PubMed abstracts (stopwords).

Discussion

Style guides, including Prescription, advise using alternative, usually simpler words where there is a choice. It was in these comparisons that I was most interested. Because the use of a word in a medical article may be different from its use in fiction, some comparisons are not of strict synonyms, but the term will do.

I am not going to give here the reasons why, for example, given is almost always a better choice of word than *administered*; the reasons for avoiding "pseudodignified prose" can be found in Prescription, and in many other style guides. What these figures provide is some evidence for which errors of style – ie, choice of word – are most prevalent, how medical English is changing, and whether those changes are internal – the influences existing within medicine and medical writing; or external – the influences coming, at least in part, from more general English. I suspect also that medical English is a sample of technical English in general, as Pinker describes [5].

As long as the exact figures for prevalence are not looked at too closely, and not compared too finely, I believe the figures to be valid. For less common words, increases in prevalence may be overestimated if the average number of words in an abstract was greater in 2010 than in 1975. I did not measure that number precisely, but a random file of 200 PubMed abstracts published in 1975 contained ~35,000 words, and a similar file from 2010 contained ~47,000. At most, by this rough and ready estimate, abstract length has increased about 1.3-fold; but these files included titles, authors, and institutions, and there is no doubt that the number of authors per article is increasing[6]. I think it unlikely that such a small increase in length of abstract can be an important factor in the many-fold increases in word prevalences.

It must be added that I have considered only simple choice of word: an aspect of style that can be studied quantitatively. There are other aspects of word choice, for example, tautology, ambiguity and cliché, and other areas of usage, for example, punctuation, concordance and tenses, which would be difficult to measure, and even more difficult to compare quantitatively between medical and more general English.

PubMed constantly accumulates articles, including from past years, and Google searches more books. The numbers reported here will change. I almost certainly made some transcription errors, but I am certain that these will affect no more than a few words, and that there are no systematic errors altering my general conclusions.

Almost all the up-words have entries in Prescription in which we give advice. Six of the words in the first ten are also up-words on Ngram, from which I conclude that the reason their prevalences have increased includes factors from outside medical English. The increased prevalence of *gender* is probably partly due to gender politics and gender assignment, rather than simply to the (usually incorrect) substitution for *sex*. *Novel*[7] increased 18-fold and now occurs in 6% of all abstracts (in 2010: for 2014 it is 8.5%), but there has been no influence from more general English. The verb *address*[8] shows no increase on Ngram: the infinitive *to address* (which cannot be separated from the present indicative in PubMed) increased a modest 1.5-fold. The high placing of forms of *impact* was expected: *impact* is replacing *effect*, and *impacts* is replacing *affects* [see[4] p. 21], and these changes have occurred also on Ngram, especially for the verb. The other up-words on Ngram were *option*, *optimization*, *mitigate*, and *randomization*. It is no surprise that the prevalence of *randomiz/sed* has increased on PubMed with the large increase in randomized controlled trials that has occurred between 1975 and 2010 (~1000 in 1975 and ~22,500 in 2010), but its prevalence has also increased in English fiction: perhaps novelists have started writing about randomized controlled trials.

The up-word *target* also warrants mention. Overall, it is a sevenfold up-word, but that masks the very large increase in prevalence of the participles: *targeting*, at 406-fold, was the largest up-word I found, with *targeted*, at 127-fold, the second largest. From a scan of abstracts, the main reason is the increasing study of targeted cellular receptors and use of targeted medical treatments, but that cannot be all. The verb form of *target* was a 6-fold up-word on Ngram: *target* is becoming like *involve* and *address*. They are catch-all words, idly chosen when the writer hasn't thought about what might be the correct word. *Target* is the perfect catch-all, because it can be used as noun, verb and adjective: *Our target is to target resources to the target population*, rather than, *We need to give money to the right people*.

Many of the down-words – *basically* and *essentially*, *former* and *latter*, *noted*, *attained* and *concerned*, etc. – we counsel against, but I don't suppose their decreasing prevalence is because of style guides, otherwise the prevalence of *plethora* wouldn't have increased 8-fold.

The results in Table 2 are the clearest evidence for pseudodignification. For most synonym pairs, the "disapproved" is relatively commoner than the "approved" in PubMed than on Ngram. Take the simple word *showed* (Table 2, lower): the medical writer is four times more likely than the writer of English fiction to write *demonstrated*, twice as likely to write *exhibited*, and one and a half times more likely to write *revealed*. There are times when the longer words are appropriate, but all those with an interest in medical prose know they are over-used.

One of the hallmarks of medical writing is *administered* for *given* (Table 2, upper, line 6). In 2010, it was nearly 40 times more likely in medical than more general English, which had increased from 25 times in 1975, and from only eight times in 1930. These increases are so for most of the comparisons in the Table: in 1930, the choice of words in medical writing was more similar to fiction.

Some of the numbers in Table 2 that don't follow the general pattern may be because of small numbers distorting interpretation: for example, there was only one *impact* in PubMed 1921-1940 (there were 908 *effect*), and *optimal* was an extremely uncommon word on Ngram in 1930.

I cannot say whether pseudodignification is due to native English speakers, or to writers for whom English is an additional language, but they must be a factor: in 1975, 68% of PubMed articles were written in English; in 2010, it was 94%. However, in 1921-40, only seven of the more than 90,000 articles were not in English, and the opacity of medical English pre-dates that: the *Lancet*[9] [and see[4] p. vi] noted in 1885 "the selection of the very longest and most technical words which the medical vocabulary will supply. This is an error to be deplored and reprobated."

There is evidence that the tendency of medical writers to choose less simple words is becoming greater. Limited evidence from 80 years ago suggests that the process of pseudodignification started a long time ago, but it seems to be accelerating. As long as there is no benefit to medical writers to write more clearly, it is difficult to see how the process can be reversed, but we must carry on trying if we are not to be lost in a polysyllabic fog.

Acknowledgements
I am grateful to Paul Gelling, Elise Langdon-Neuner and Adam Jacobs for helpful comments and suggestions.

References

1 O'Donnell M. One man's burden. *BMJ* 1985;290: 250.
2 http://www.ncbi.nlm.nih.gov/pubmed (accessed many times)
3 https://books.google.com/ngrams (accessed many times)
4 Goodman NW, Edwards MB, Langdon-Neuner E. *Medical writing: a prescription for clarity*. 4th edition. Cambridge: CUP, 2014.
5 Pinker S. The curse of knowledge in *The sense of style*. London: Allen Lane, 2014, pp. 57-76.
6 King C. Multiauthor papers: onward and upward. http://archive.sciencewatch.com/newsletter/2012/201207/multiauthor_papers/ (accessed 11 February 2015).
7 Goodman NW. Paradigm, parameter, paralysis of mind. *BMJ* 1993;307;1627-1629.
8 Goodman NW. Addressing issues or writing what you mean. *The Write Stuff* 2008;17(4):181-182.
9 Anon. From the Lancet. *The Lancet* 1990; 336:224.

New EASE secretary

EASE secretary Tina Wheeler will be stepping down in May 2015 and will be replaced by Dalibora Behmen. Tina joined the EASE secretariat in April 2012 and has calmly and efficiently organised everything since. Many EASE members will have met Tina during the conferences in Split or Blankenberge, and others will have dealt with her by email or phone. As well as overseeing the daily running of EASE and supporting Council, Tina helped with a major upgrade of the membership database.

> "EASE members are such a nice bunch! I have enjoyed my time as EASE secretary and was especially pleased to meet up with some members at conferences. Being the EASE secretary has been a privilege and I wish Dalibora every success."

Everyone at EASE would like to send their best wishes for Tina's well-earned retirement, much of which will doubtless be spent with her beloved horse, Echo.

Dalibora is the Head of Office at the Croatian Centre for Global Health, University of Split School of Medicine. She has a degree in English and Italian languages and literature, and a master's degree in business economics. Her thesis explored the functioning of management in institutions of higher education. She has a strong interest in project administration and has attended numerous national and international workshops related to this topic, with special emphasis on education in project management and finance. She is experienced in organizing events and is very skilled in coordinating all types of activities.