

A Statistical Analysis of the Vocabulary of Medical Research Articles (2): Differences across the “IMRAD” Structure

Faculty of Social Information Science, Kure University
Simon A. Fraser

Abstract This is the second in a series of three papers attempting to quantify the vocabulary of medical research articles. It continues my investigation into a word frequency count derived from a 21,000 word computer corpus of articles drawn from a range of specialized medical fields. The distribution of the most frequent words across the “IMRAD” structure of medical articles is investigated. Several words are found to be distributed differently across the divisions, and it is shown that the way in which a word is used is often dependent on the section in which it is found. These findings have implications not only for teachers and learners of Medical English, but also for those in the wider field of ESP (English for Specific Purposes).

Key Words: Medical English, Computer corpus, Frequency count, Word distribution, “IMRAD” structure

1 INTRODUCTION

This paper is the second in a series of three articles comprising a statistical analysis of the vocabulary of medical research articles. The papers investigate a frequency count derived from a computer corpus representing a sample of language used in the medical research article. In the first paper (Fraser¹⁾), a comparison was made between the most frequent items on the medical list and those on the Cobuild corpus of general English. The frequency and distribution of lexical items were found to differ considerably between the two lists. It was suggested that a general frequency list is therefore inadequate as a source of vocabulary selection for a medical English curriculum, and that priority should be given to the most frequent words with high distribution across the different medical fields.

The medical research article is commonly divided into four standardized divisions, each

representing a broad rhetorical function: Introductions (I), Methods (M), Results (R), and Discussions (D) – henceforth known as “IMRAD”. Here, I would like to examine the way in which the most frequent words are distributed across IMRAD, and look in some detail at those words with interesting distribution.

2 THE ANALYSIS

For the statistical analysis I used the corpus of sixteen medical articles (21,000 words) compiled at the Institute for Applied Language Studies, Edinburgh University. The articles covered a wide range of medical fields, and were taken from both British and American journals. (Appendix 1 lists the articles in the IMRAD corpus. For a more detailed description of the analysis, its limitations, and the frequency word list produced by the computer, see Fraser¹⁾).

3 WORDS WITH INTERESTING DISTRIBUTION ACROSS "IMRAD"

After examination of the medical word list produced by the computer, the following words were found to be distributed differently across the different rhetorical divisions. In the tables, the first column (Freq.) shows the actual frequency of occurrence in the corpus, both as a "raw" figure and expressed as a percentage of occurrence in the entire article. The figures in the second row of each entry show the occurrence of the word expressed as a percentage of the total number of words in a particular section.

3.1 Was and Were

TABLE 1

	I	M	R	D	Total
was Freq	7 (2.4%)	134 (46.5%)	84 (29.2%)	63 (21.9%)	288 (100%)
%	0.24	2.11	1.93	0.77	1.33
were Freq	8 (3.0%)	133 (50.6%)	67 (25.5%)	55 (20.9%)	263 (100%)
%	0.28	2.09	1.54	0.68	1.21

Both of these word forms occur far more frequently in Methods and Results than in the other sections. This is unsurprising, as we would expect the procedural Methods and Results sections to favor the past tense, and is in agreement with Heslot's²⁾ figures of 94% for the simple past tense in these sections. That *was* and *were* are most frequent in Methods can be explained by the fact that this section strongly favors passive constructions such as the following:

1. Statistical analysis *was* done with the McNemar's test for matched samples.

(Paper 14, p.14, M)

3.2 Is, Are and Be

TABLE 2

	I	M	R	D	Total
is Freq	40 (21.5%)	22 (11.8%)	8 (4.3%)	116 (62.4%)	186 (100%)
%	1.39	0.35	0.18	1.42	0.86
are Freq	18 (22.5%)	8 (10.0%)	6 (7.5%)	48 (60.0%)	80 (100%)
%	0.62	0.13	0.14	0.59	0.37
be Freq	20 (17.4%)	10 (8.7%)	6 (5.2%)	79 (68.7%)	115 (100%)
%	0.69	0.16	0.14	0.97	0.53

These forms show similar distribution across IMRAD. They are all far higher in Introductions and Discussions than in the other two sections. The high frequency of *is* and *are* suggests that the present tense occurs principally in these two sections, which concurs with the findings of Heslot²⁾. The following examples show how *is*, *are* and *be* are used:

1. Typhoid fever *is* a serious disease occurring frequently in many developing countries.
(Paper 1, p.1101, I)
2. Such incidents *are* especially likely if surveillance in the home is compromised by...
(Paper 12, p.521, D)
3. Improvement in theoretical knowledge, however, may *be* greater than that in behavior at the roadside.
(Paper 7, p.1196, D)
4. The environmental factor could *be* genetic inbreeding...
(Paper 5, p.752, D)

Examples 1 and 2 show that the present tense typically has the rhetorical function of expressing generalizations and conclusions based on the results of research. We would therefore expect *it*, and thus *is* and *are*, to occur most frequently in the Introduction and Discussion sections. Examples 3 and 4 show how *be* is commonly used — in modal expressions. Modals are also found to occur mostly in Introductions and Discussions (see Table 4).

3.3 Have, Has and Been

TABLE 3

	I	M	R	D	Total
have Freq	19 (25.0%)	8 (10.5%)	7 (9.2%)	42 (55.3%)	76 (100%)
%	0.66	0.13	0.16	0.52	0.35
has Freq	27 (49.1%)	7 (12.7%)	1 (1.8%)	20 (36.4%)	55 (100%)
%	0.94	0.11	0.02	0.25	0.25
been Freq	22 (32.3%)	15 (22.1%)	8 (11.8%)	23 (33.8%)	68 (100%)
%	0.76	0.24	0.18	0.28	0.31

All of these words occur with especially high frequency in Introductions. If we look at the research articles for an explanation, we see that a large number of them use the present perfect tense to introduce the reader to a specific area and to focus on the current state of

affairs. Many of these sentences use *has+been* or *have+been*, as in the following examples:

1. The immune response to influenza virus infection *has been* studied extensively (see review by Ada and Jones).

(Paper 6, p.477, I)

2. Despite the relative frequency of such cases there *have been* few reports on the problem...

(Paper 12, p.519, I)

3. In recent years, food intolerance *has been* suggested as a causative factor in numerous childhood conditions...

(Paper 2, p.1696, I)

3.4 Modals

TABLE 4

		I	M	R	D	Total
may	Freq	8(17.8%)	2 (4.4%)	0 (0%)	35 (77.8%)	45 (100%)
	%	0.28	0.03	0	0.43	0.21
would	Freq	2 (7.7%)	4 (15.4%)	1 (3.8%)	19 (73.1%)	26 (100%)
	%	0.07	0.11	0.02	0.23	0.12
can	Freq	7(31.8%)	0 (0%)	1 (4.5%)	14 (63.7%)	22 (100%)
	%	0.24	0	0.02	0.17	0.10
could	Freq	1 (4.8%)	5 (22.7%)	1 (5.9%)	12 (70.6%)	17 (100%)
	%	0.03	0.08	0.02	0.17	0.10
will	Freq	1 (5.9%)	3 (17.6%)	1 (5.9%)	12 (70.6%)	17 (100%)
	%	0.03	0.05	0.02	0.15	0.08
should	Freq	1 (7.1%)	0 (0%)	0 (0%)	13 (92.9%)	14 (100%)
	%	0.03	0	0	0.16	0.06

3.4.1 May

May occurs with the highest frequency in the Discussion section of medical articles. It is not frequent in Methods, and it does not occur at all in Results. This finding provides support for the conclusions of Adams Smith³⁾, who investigated "author's comment" in six medical research papers. She found that authorial comment, which is mainly introduced by modal auxiliaries, is far more common in Introductions and Discussions than in the other two sections. Most of the authorial comment in the six articles is epistemic, i.e. relating to the writer's knowledge of the situation in question. In Fraser¹⁾, it was found that *may* is mainly used to express epistemic modality. We would therefore expect *may* to be particularly frequent in the Introduction and Discussion sections of medical research articles.

3.4.2 Would

Would follows a similar distribution pattern to that of *may*, but occurs with much higher frequency in Discussions than in Introductions. This is perhaps because it is often used to make recommendations which are based on the results of the research. For example:

...it *would* therefore be sensible not only to do the customary neurological investigations but also to search for the risk factors for stroke...

(Paper 14, p.12, D)

3.4.3 Should

Should is also found to the greatest extent in the Discussion section. It is often used here, in its deontic sense, to make suggestions or to express obligation resulting from the research findings:

We recommend, therefore, that more patients *should* be offered exposure to positive suggestions...

(Paper 4, p.789, D)

3.4.4 Will

The distribution of *will* across IMRAD is very similar to that of *should*. It seems to be used particularly to make future predictions, so we would expect it to occur with high frequency in Discussions.

Although this approach *will* have major financial and logistic consequences...

(Paper 13, p.551, D)

3.4.5 Can

Can occurs with the highest frequency in the Introduction section, followed by the Discussion section. It is the only modal to be found less frequently in Discussions. This may be because *can* is not used as often as, for example, *may* or *could* when making predictions based on the results of the author's research.

3.4.6 Could

Could occurs with a much higher frequency

in Discussions than in any other section.

Presumably this is because, along with *may*, it is often used when the author does not want to state her views too definitely. It is interesting to note that although *could* is frequent in Methods, *can* does not occur in this section at all. This fits in with the observation that in the Methods section, the past tense is used almost exclusively.

3.5 We

TABLE 5

	I	M	R	D	Total
we Freq	9 (20.4%)	8 (18.2%)	0 (0%)	27 (61.4%)	44 (100%)
%	0.31	0.13	0	0.33	0.20

We is found mainly in the Introduction and Discussion Sections, and to a lesser extent in Methods. It does not occur at all in Results. This accords with Adams Smith's⁴⁾ observation that authorial comment is high in Introductions and Discussions and very low in Methods and Results.

3.6 If

TABLE 6

	I	M	R	D	Total
if Freq	1 (2.8%)	11 (30.6%)	1 (2.8%)	23 (63.9%)	36 (100%)
%	0.03	0.17	0.02	0.28	0.17

This word occurs with the highest frequency in the Discussion section, followed by Methods. Its occurrence is surprisingly low in Introductions. It is surprising that *if* occurs with such high frequency in Methods. The following examples show how it is commonly used in that section:

1. Patients were considered to have acute appendicitis *if* the skin temperature on the right was at least 1°C warmer than on the left. (Paper 8, p.722, M)
2. *If* the clinician then deemed that a barium enema was required that patient was included in the protocol. (Paper 13, p.549, M)

We can see that *if* is used to indicate that there is some prerequisite before the next step

of a procedure can be carried out, or a diagnosis can be made, for example.

3.7 However

TABLE 7

	I	M	R	D	Total
however Freq	8 (25.0%)	0 (0%)	4 (12.5%)	20 (62.5%)	32 (100%)
%	0.28	0	0.09	0.25	0.15

Unsurprisingly, *however* occurs most often in Introductions and Discussions. There are no instances of it at all in the Methods section. It was found in Fraser¹⁾ that *however* is used to signal the introduction of a problem, so we would expect its frequency to be particularly high in I and D. In I, *however* is used to indicate a gap in research, and in D, it draws attention to the problems which still remain, further work which needs to be done, and suggestions which need to be made.

3.8 Therefore

TABLE 8

	I	M	R	D	Total
therefore Freq	3 (25.0%)	0 (0%)	0 (0%)	9 (75.0%)	12 (100%)
%	0.10	0	0	0.11	0.06

The distribution of this word is similar to that of *however*. It is equally distributed across Introductions and Discussions, and there are no occurrences in Methods or Results. We would expect *however* and *therefore*, which are discourse markers, to occur with the highest frequency in I and D.

3.9 Thus

TABLE 9

	I	M	R	D	Total
thus Freq	1 (5.6%)	3 (16.7%)	3 (16.7%)	11 (61.1%)	18 (100%)
%	0.03	0.05	0.07	0.14	0.08

Thus, as we might expect, is particularly frequent in the Discussion section, but it is interesting that it is found least in Introductions. This would suggest that *thus* is used primarily by the author when describing or discussing his *own* results, rather than those of previous research.

3.10 *Since*

TABLE 10

	I	M	R	D	Total
<i>since</i> Freq	2 (10.5%)	4 (21.1%)	0 (0%)	13 (68.4%)	19 (100%)
%	0.07	0.06	0	0.16	0.09

Since occurs with a particularly high frequency in Discussions, where it is found more often than in the other sections put together. The Discussion section has the function of explaining the findings, and *since* is often used, as in the following example:

Multiple colonic biopsies are recommended... *since* 5 patients whose colons were normal were subsequently found. (Paper 13, p.551, D)

3.11 *Table*

TABLE 11

	I	M	R	D	Total
<i>table</i> Freq	0 (0%)	5 (16.1%)	26 (83.9%)	0 (0%)	31 (100%)
%	0	0.08	0.60	0	0.14

It is not at all unexpected to find *table* only in Methods and Results. The vast majority of these occurrences are in Results, in which findings are often presented in tabular form.

3.12 *Shows*

TABLE 12

	I	M	R	D	Total
<i>shows</i> Freq	0 (0%)	0 (0%)	11 (84.6%)	2 (15.4%)	13 (100%)
%	0	0	0.02	0.02	0.06

As expected, *shows* occurs almost exclusively in the Results section, where it is often found to collocate with *table*:

Table II shows the morphine requirements over 24 hours after the operation...

(Paper 4, p.789, R)

3.13 *Reported*

TABLE 13

	I	M	R	D	Total
<i>reported</i> Freq	6 (23.1%)	4 (15.4%)	3 (11.5%)	13 (50.0%)	26 (100%)
%	0.21	0.06	0.07	0.16	0.12

As we would expect, *reported* is found mainly in the Introduction and Discussion sections. In Introductions, it is generally used when introducing past research, and the present perfect tense is often used (see Example 1). In Discussions, it occurs most frequently when other researchers' work is being compared with the author's own study (Example 2).

1. Epilepsy following a clinically or pathologically recognizable stroke has been widely *reported*. (Paper 14, p.11, I)
2. Other workers have not *reported* how patients were monitored. (Paper 10, p.258, D)

3.14 *Such*

TABLE 14

	I	M	R	D	Total
<i>such</i> Freq	10 (38.5%)	4 (15.4%)	0 (0%)	12 (46.1%)	26 (100%)
%	0.35	0.06	0	0.15	0.12

Such is particularly frequent in Discussions, and the following examples show how it is typically used:

1. ...with other histological variants *such as* mesangial proliferative glomerulonephritis. (Paper 10, p.255, I)
2. Despite the relative frequency of *such* cases there have been few reports on the problem... (Paper 12, p.519, I)

It is often found in the phrase *such as*, which is used for exemplification of previous findings (Example 1). It is also used anaphorically, to refer back to a previous stretch of text (Example 2).

3.15 *Then*

TABLE 15

	I	M	R	D	Total
<i>then</i> Freq	2 (11.8%)	14 (82.3%)	0 (0%)	1 (5.9%)	17 (100%)
%	0.07	0.22	0	0.01	0.08

Nearly all occurrences of this word are in the first two sections of the articles, with most of them in Methods. This should not be surprising

when we consider that it is in this section that the experimental procedure is being described and explained. We can see that *then* is often used in Methods when the steps of a process are being outlined:

With the patient remaining in the prone position, the skin of the calf was *then* cleaned with antiseptic. (Paper 11, p.1099, M)

3.16 They

TABLE 16

	I	M	R	D	Total
they	1 (5.9%)	10 (58.8%)	6 (35.3%)	0 (0%)	17 (100%)
	0.03	0.16	0.11	0	0.08

They occurs almost exclusively in Methods and Results, which implies that this pronoun is only used to refer to patients or experimental subjects, and rarely used to refer to other researchers, for example.

3.17 Received, Taken, Asked and Done

TABLE 17

	I	M	R	D	Total
received	0 (0%)	11 (73.3%)	3 (20.0%)	1 (6.7%)	15 (100%)
	0	0.17	0.07	0.01	0.07
taken	1 (6.7%)	10 (66.7%)	2 (13.3%)	2 (13.3%)	15 (100%)
	0.03	0.16	0.05	0.02	0.07
asked	0 (0%)	10 (100%)	0 (0%)	0 (0%)	10 (100%)
	0	0.16	0	0	0.05
done	1 (10.0%)	8 (80.0%)	1 (10.0%)	0 (0%)	10 (100%)
	0.03	0.13	0.05	0	0.05

These words all occur in Methods far more frequently than in any other section. As they are all "past" forms of verbs, we would expect them to occur especially in Methods and Results. They are particularly frequent in Methods because of the preponderance of the passive voice in that section.

3.18 Through

TABLE 18

	I	M	R	D	Total
through	1 (7.1%)	9 (64.3%)	1 (7.1%)	3 (21.4%)	14 (100%)
	0.03	0.14	0.02	0.04	0.06

Through occurs much more frequently in Methods than in any other section. One simple reason for this is that it was used several times in the American journals in the following way, when discussing a time interval:

Monthly birth data for January, 1972, *through* December, 1980... (Paper 5, p.749, M)

3.19 While

TABLE 19

	I	M	R	D	Total
while	0 (0%)	0 (0%)	8 (57.1%)	6 (42.9%)	14 (100%)
	0	0	0.18	0.07	0.06

This word is found only in the Results and Discussion sections, with most occurrences in Results. It is often used to show what was going on at the time something of medical significance occurred, as in the following example:

Four toddlers became intoxicated *while* actually attending family celebrations.

(Paper 12, p.520, R)

3.20 Important

TABLE 20

	I	M	R	D	Total
important	2 (15.4%)	0 (0%)	1 (7.7%)	10 (76.9%)	13 (100%)
	0.07	0	0.02	0.12	0.06

Important occurs most frequently in Discussions, followed by Introductions. This is because *important* is a modifier which helps to express the author's attitudes and evaluations of the content of the text.

The recent legislation enacted by parliament concerning the use of rear seat belts for children is an *important* preventative measure...

(Paper 7, p.1196, D)

4 CONCLUSION

This paper continued my analysis of a computer corpus of 16 medical research articles

taken from seven different British and American journals. I investigated the distribution of words across the “IMRAD” (Introduction, Methods, Results, Discussion) structure of the articles. Several words were found to be distributed differently across these rhetorical divisions; some, in fact, were found exclusively in a particular section or sections.

Although the sample size is admittedly small for a study using a computer corpus, the results should be of considerable interest to teachers and learners of Medical English, and

indeed to those working in the wider field of ESP. Learners’ attention should be drawn to the “IMRAD” structure of the medical research paper. It should be pointed out that certain lexical items may be found with high frequency in particular sections of the paper because they are performing a specific rhetorical function (e.g. describing, evaluating, introducing a “research space”). It is important to realize that the way in which a word is used will often depend on the section in which the word is found.

References

- 1) S. A. Fraser. 2001. A statistical analysis of the vocabulary of medical research articles (1): comparison with the Cobuild frequency count. *Integrated Studies in Nursing Science*. 3: 38–58.
- 2) J. Heslot. 1982. Text and other indexical markers in the typology of scientific texts in English. In J. Hoedt et al. (eds). *Pragmatics and LSP*. Copenhagen: Copenhagen School of Economics.
- 3) Diana E. Adams Smith. 1984. Medical discourse: aspects of author’s comment. *The ESP Journal* 3: 25–36.

APPENDIX 1

The Medical Journal Articles

PAPER 1

I. L. Acharya, C. U. Lowe et al. 1987. Prevention of typhoid fever in Nepal with the Vi capsular polysaccharide of salmonella typhi. *The New England Journal of Medicine* 317 (18): 1101–1103.

PAPER 2

C. E. Price, R. J. Rona et al. 1988. Height of primary school children and parents’ perceptions of food intolerance. *British Medical Journal* 296: 1696–1699.

PAPER 3

H. E. Nelson, S. Thrasher, T. R. E. Barnes 1991. Practical ways of alleviating auditory hallucinations. *British Medical Journal* 302: 327.

PAPER 4

T. T. C. McLintock, H. Aitken et al. 1990. Postoperative analgesic requirements in patients exposed to positive intraoperative suggestions. *British Medical Journal* 301: 788–790.

PAPER 5

A. D. Strickland, K. Shannon. 1982. Studies in the etiology of extrahepatic biliary atresia: Time-space clustering. *The Journal of Pediatrics* 100: 749–753.

PAPER 6

F. Vacheron, A. Rudent et al. 1990. Production of interleukin 1 and tumour necrosis factor activities in bronchoalveolar washings following infection of mice by influenza virus. *Journal of General Virology* 71: 477-479.

PAPER 7

P. M. Sharples, A. Storey et al. 1990. Causes of fatal childhood accidents involving head injury in Northern region. *British Medical Journal* 301: 1193-1197.

PAPER 8

J. E. Hambidge. 1990. Use of skin thermometer to diagnose acute appendicitis. *British Medical Journal* 300: 722.

PAPER 9

G. Lamont, C. J. Simpson et al. 1986. Does cimetidine alter the prognosis following perforated duodenal ulcer? *Scottish Medical Journal* 31: 198.

PAPER 10

M. A. Lewis, E. M. Baildom et al. 1989. Nephrotic syndrome: From toddlers to twenties. *The Lancet*, February 4: 255-257.

PAPER 11

T. O' Brien. 1984. The needle test for complete rupture of the Achilles tendon. *The Journal of Bone and Joint Surgery* 66-A: 1099-1101.

PAPER 12

J. O. Beattie, D. Hull, F. Cockburn 1986. Children intoxicated by alcohol in Nottingham and Glasgow, 1973-84. *British Medical Journal* 292: 519-521.

PAPER 13

P. Durdey, P. M. T. Weston, N. S. Williams. 1987. Colonoscopy or barium enema as initial investigation of colonic disease. *The Lancet*, September 5: 549-551.

PAPER 14

R. A. Shinton, J. S. Gill et al. 1987. The frequency of epilepsy preceding stroke: Case control study in 230 patients. *The Lancet*, January 3: 11-12.

PAPER 15

Managing drug dealers who swallow the evidence. *British Medical Journal* (authors and date unknown).

PAPER 16

Snoring as a risk factor for ischaemic heart disease and stroke in men. 1985. *British Medical Journal*, September 7 (authors unknown).