Navigating the challenges of medical English education: a novel approach using computational linguistics

Alberto Alexander Gayle[1,2]

[1]Center for Medical and Nursing Education, Mie University School of Medicine, Mie, Japan

[2]Department of Immunology, Mie University Graduate School of Medicine, Mie, Japan

* Corresponding author

E-mail: aruberutou@doc.medic.mie-u.ac.jp (AG)

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1711v2 | CC-BY 4.0 Open Access | rec: 6 Feb 2016, publ: 6 Feb 2016

Abstract

Recent studies have shown that international medical graduates (IMG) comprise a substantial and increasingly larger share of the medical workforce, internationally. IMGs wishing to work in English-speaking countries face many challenges. And overcoming such challenges plays an important role in ensuring a more comfortable transition and improved outcomes for patients. This study addresses one such area of concern: the efficient acquisition of advanced language competence for use in the medical workplace. This research also addresses the needs of medical students and practitioners in other countries, where English is not the primary language.

Medical terminology and phrasing is based on a tradition spanning more than 2500 years—a tradition that cuts across typical linguistic and cultural boundaries. Indeed, as is commonly understood, the language required by doctors and other medical professionals varies substantially from the norm. In the present study, this dynamic is exploited to identify and characterize the language and patterns of usage specific to medical English, as it is used in practice and reporting.

Overall, constructions comprised of preposition-dependent nouns, verbs and adjectives were found to be most prevalent (38%), followed by prepositional phrases (33%). The former includes constructions such as “present with”, “present to”, and “present in”; while constructions such as “of … patient”, “in … group”, and “with … disease” comprise the latter. Preposition-independent noun and verb-based constructions were far less prevalent overall (18% and 5%, respectively).

Up to now, medical language reference and learning material has focused on relatively uncommon, but essential, Greek and Latin terminology. This research challenges this convention, by demonstrating that medical language fluency would be acquired more efficiently by focusing on prepositional phrases or preposition-dependent verbs, nouns, and adjectives in context. This work

23 should be of high interest to anyone interested in improved communication competence within the

24 English-speaking medical workplace and beyond.

What this paper adds
What is already known on this subject
• International medical graduates make up a substantial portion of the medical workforce • Imperfect medical English creates challenges for international medical graduates • Subideal language impacts credibility and has been associated with increased risk to patients
What this study adds
• Preposition-dependent terms, following Germanic usage patterns, dominate medical English • Complex terms derived from Greek and Latin are far less prevalent than assumed • Medical English learning expected to be expedited by focus on preposition-dependent terms

25

26

27

28

Introduction

International medical graduates (IMGs) have become essential to the health care systems of much of the developed world. [1] According to the latest OECD figures, IMGs represent 17.6% of the medical workforce among the OECD26 nations. [2] This figure is even higher within the English-speaking world, with United States, United Kingdom, Canada, and Australia each reporting well over 20% (25.0%, 28.7%, 23.5%, and 30.5%; respectively). IMGs often face many challenges when entering the workforce. These challenges have been well documented [3–5] and are often boiled-down to issues requiring improved acclimatization and communication. [6] Language barriers have often been cited as a key challenge [7], but very few studies have been conducted to explore what specific content or grammatical features are omitted by IMGs in practice. [8]

Beyond the clinical setting, many have claimed that the ubiquity of English in medical reporting and communications has created undue burden for non-native speakers of English (nNS). [9–11] And the evidence supports these claims: nNS clinician-scientists are more likely to have had their research rejected for publication [9] or retracted due to reporting misconduct. [12] Not surprisingly, they are also reported to be overall less satisfied with the publication process. [13]

The history and intellectual sophistication encoded within the language of medicine makes tackling such issues no easy feat. Indeed, medical terminology and phrasing are based on a tradition spanning more than 2500 years—a tradition that cuts across typical linguistic and cultural boundaries. [14] Consequently, the phrasing and language patterns typical within medical English vary substantially from that which would be considered typical of the common language. [15] This dynamic, however, is exploited in the present study to uncover and characterize the specific content and language patterns most prevalent within English, as it is used within medicine.

Methods

To accomplish these goals, methods from the area of computational linguistics were applied to a representative language database (corpus) in order to extract and derive the following information: 1) the collocations most likely to appear in medical English writing, 2) the proportion of the various parts of speech (and associated phrases) present in medical English writing, and 3) examples of representative language for each. In order to accomplish this, it was necessary to obtain a corpus that was sufficiently representative with respect to medical English. For this purpose, we used the Oxford English Corpus (OEC). The OEC is is one of the largest English-language corpora in the world, and is used by the Oxford University Press to support the production of their famed series of dictionaries of the English language and associated material. As stated by the Oxford University Press, "the corpus contains nearly 2.5 billion words of real 21st century English". [16] It is considered to be the largest structured corpus of any language. [17] In addition to full-text, each document includes the following metadata, where available: title, author, author gender, dialect, date, and subject domain. In addition, document statistics (word count, sentence count, et al.) are generated automatically for each.

All data preparation and analysis was conducted using the SketchEngine corpus management and analysis software by Lexical Computing Limited (UK). [18] SketchEngine provides many tools for analyzing the relationships between words within and across documents. This includes sophisticated analyses of collocation that enable the construction of statistically generated thesauri, concordances, and comparative analyses. [19] Downloading and processing of final data was done using Rapidminer Studio 6.5 (RapidMiner GmbH. Released 2015. RapidMiner Studio Academia, Version 6.5002) with a custom script written in R.

The corpus

In order to produce a representative corpus specific to medical English, we identified and filtered all documents within the OEC classified under the subdomain "medicine". In total, this included almost 75 million tokens (74,903,294): 3.08% of the OEC total. Grammar relationships for each are calculated and assigned based on a modified version of the Penn Treebank. [20]

Term identification and ranking

Collocations were calculated within the SketchEngine using the log-likelihood algorithm described by Dunning. [21] The frequency of each collocation within the medical subcorpus was then compared to the respective frequencies within the original, full OEC corpus. The following formula was used: $score = \frac{1+freq_{sub}}{1+freq_{OEC}}$, where *sub* indicates the medical subcorpus, *OEC* indicates the full OEC corpus, and *freq* is the frequency of a given collocation within each respective corpus. [22] Collocations were then ranked according to *score*. The result was a ranked list of terms and phrases most likely to be found primarily in the medical subcorpus, along with POS information and frequency statistics. From this list, the top 10,000 collocations were retained for further analysis. This data set was downloaded using Rapidminer and converted into an Excel file. Within Excel, this list was then processed to remove duplicate entries. Following this process only 5436 entries remained. Collocations were categorized according to grammatical relationship. The average and aggregate frequencies were then calculated for each grammatical category.

Concordance generation

These remaining entries were then fed back into the SketchEngine for concordance generation. For each selected collocation, a concordance was derived from entries within the medical

subcorpus. Corpus entries were selected from among the hundreds of options by applying the "good dictionary example" (GDEX) heuristic. [23] This algorithm automatically ranks concordance entries based primarily on simplicity of structure and typicality, with respect to other potential examples. The top five concordance entries were then returned accordingly. Concordance entries were then merged with the original dataset, to produce a final table consisting of: overall frequency, primary POS, collocation, grammar relationship, and five dictionary examples (Table 1).

Table 1. Example output.

FREQ	LEFT	GR	RIGHT	Example Usage in Context
60.66	patient	N PREP	with	While the prognosis is quite good , [[patients with]] peripheral arterial occlusive
60.66	patient	N PREP	with	called by the emergency room to see a [[patient with]] possible myocardial
60.66	patient	N PREP	with	We provided a report about a [[patient with]] angiosarcoma on her scalp
50.69	of	PREP N	patient	many involved in the care of each [[of his critically ill patients]] , the patients
50.69	of	PREP N	patient	improving the accuracy in this subgroup [[of smear-negative patients]]
50.69	of	PREP N	patient	Yet until recently the wisdom and experience [[of the patient]] has been only
50.69	of	PREP N	patient	We can deduce this from the dreams [[of patients]] in analysis which
50.69	of	PREP N	patient	diagnostic techniques improved the outcome [[of patients]] with suspected
46.39	in	PREP N	patient	the low prevalence of the disease [[in referred patients]] without osteoporosis
46.39	in	PREP N	patient	therapy has been shown to be beneficial [[in patients]] with ataxia with
46.39	in	PREP N	patient	found to correlate with exercise capacity [[in patients]] with COPD , supporting
46.39	in	PREP N	patient	We had interpretable results [[in 44 patients]] as follows :
46.39	in	PREP N	patient	in future randomized trials of anticoagulation [[in cardiomyopathy patients]] .
44.07	number	N PREP	of	Zinc-deficient diets markedly increased the [[number of]] tumours generated
44.07	number	N PREP	of	Within the 29 patients with cough , the [[number of]] TRPV - 1positive
44.07	number	N PREP	of	With the same [[number of]] patients in each group
41.53	associate	V PREP	with	and low gamma-GCS reactivity may be [[associated with]] the high sensitivity
41.53	associate	V PREP	with	was effective in some patients with ataxia [[associated with]] CoQ10 deficiency
41.53	associate	V PREP	with	Weight gain was [[associated with]] increased energy , a
41.53	associate	V PREP	with	This was not [[associated with]] a significantly

Table 1. Example output. Final output generated as shown here. Each column denotes: a) frequency per million (FREQ); b) term part at the head of a given collocation (LEFT); c) term part at the end of a collocation (RIGHT); d) grammar relationship between LEFT and RIGHT (GR).

Results

Our resulting dataset consisted of 5436 collocations corresponding to the terms and phrases most likely to be found primarily within the medical English corpus. Collocations were then categorized according to grammatical relationship and frequency statistics. For the purpose of this analysis, "prevalence" is defined as the number of terms per category multiplied the average frequency; "prevalence" may also be derived by summing up the respective frequencies of each term within a given class. Prevalence thus reflects the relative likelihood that a given class of grammar relationship might be encountered within a given medical text. For the remainder of this paper, prevalence is described as a percentage with base of n = 9561. Table 2 summarizes these findings.

Table 2. Top grammar relationships within medical English usage.

Grammar	Description	Term Count (Absolute)	Term Count (Percent)	Term Freq. (Average)	Prevalence (Absolute)	Prevalence (Percent)
PREP N	Noun introduced by preposition	1535	28.2%	2.1	3202	33.1%
N PREP	Noun followed by preposition	944	17.4%	2.9	2721	28.1%
X mod N	Modified noun	1485	27.3%	1.1	1624	16.8%
V PREP	Verb followed by preposition	286	5.3%	2.3	661	6.8%
and/or	Coordinated pair	315	5.8%	1.0	328	3.4%
V obj N	Verb-object collocation	281	5.2%	1.2	327	3.4%
ADJ PREP	Adjective-preposition pairs	88	1.6%	2.8	247	2.6%
X of N	"of" modified noun	153	2.8%	1.0	157	1.6%
ADV V	Common modified verbs II	42	0.8%	1.2	51	0.5%
X to N	"to" Noun	23	0.4%	2.1	49	0.5%
V Part	Separable verbs	20	0.4%	2.0	41	0.4%
ADV ADJ	Modified or intensified adjectives	28	0.5%	1.3	37	0.4%
N subj V	Subject-verb collocation	32	0.6%	1.1	34	0.4%
N is ADJ	Common "is" expressions	16	0.3%	1.9	30	0.3%
X to V	Common infinitives with "to"	15	0.3%	1.0	14	0.1%
V ADV	Common modified verbs I	9	0.2%	1.4	12	0.1%
it+	Common "it is" expressions	6	0.1%	1.9	12	0.1%
X than N	Comparative using "than"	5	0.1%	1.0	5	0.1%
		** Average term frequency less than 1.0 omitted.*				
		** term count less than 5 omitted.*				

Table 2. Top grammar relationships within medical English usage. Term Count (Percent) includes terms not shown; n = 5436. Term Frequency reflects term occurrence per million terms. Prevalence (Percent) is based on overall aggregate prevalence (count * frequency), n = 9561.

Prevalence within medical English

The most prevalent types of terms were nouns introduced by prepositions ("Prep N", 33%), nouns followed by prepositions ("N Prep", 28%), and modified nouns ("X mod N", 17%); verbs followed by prepositions ("V Prep", 7%) came in a distant fourth. The long-tail consisted of another 30 miscellaneous grammar relationships accounting for 15% of total prevalence.

"Prep N" includes collocations such as "of … patient", examples of which include the expressions "of this critically ill patient", "of the patient", and "of patients". Some other high-frequency collocations were "in ... group" (e.g., "in the control group" and "in both groups") and "of study" (e.g., "of this study" and "of the previous studies"). "N Prep", captured nouns for which appropriate collocation usage and/or meaning depends on the preposition. This includes terminology such as "treatment of", "treatment in", and "treatment with". The third category, "X mod N" was found to be comprised mostly of noun-adjective pairs forming specific terminology such as "blood pressure", "risk factor", "side effect", and "heart disease". The fourth place category, "V Prep", captured verbs for which the meaning conveyed depends on the preposition used. This includes terms such as "present with", "present to", "present in", and "present as".

The remaining 15% of grammatical relationships observed are listed in Table 2. In total, thirty classes of grammatical relationships were observed. Of these, however, only a few stood out. The rest comprised a substantial long-tail of potentially informative, but low incidence terms and phrases; these were omitted from analysis. A brief overview of the remaining, significant terms follows.

"And/Or" (3%) is comprised mostly of common noun and modifier pairs (coordinates) that are generally equivalent, either in terms of collocational usage or meaning (modifiers). Examples of

noun coordination include “children and adolescents”, “drugs and alcohol”, and “anxiety and depression”. Examples of modifier coordination include “negative and positive”, “male and female”, “safe and effective” and even quasi-equivalent pairs such as “many other”.

“V obj N” (3%) captured noun-verb collocations required to appropriately describe concepts or actions specific to medical practice and/or science. Examples include “have … effect”, “treat … patient”, “provide … information”, “reduce … risk”, and “make … decision”.

“Adj Prep” (3%) represents adjectives followed by prepositions and includes items such as “due to”, “effective in”, “available for”, “aware of”, and “common in”. Items in this class generally serve as the objects of passive constructions, for example “Is due to” and “Is effective for”.

Meanwhile, “X of N” (2%) captured modified nouns inverted using the “of” construction; examples include “% … of … patients”, “quality of life”, “result … of … study”, and “cause of … death”. Many were phrase equivalent reformulations of terms captured under “X mod N”.

Discussion

In the present study, we used computational linguistic methods to systematically explore medical English as an entity separate to and apart from the English language itself. The methodology demonstrated in this study compared two separate, but non-independent corpora—one representative of general English usage, the other specific to medicine—to identify the usage patterns specific to medicine, that are less likely to be encountered in day-to-day usage.

Pedagogical implications

As previously described, grammatical relationships were identified and explored based on the findings of a computational analysis. Such approaches are known to identify and describe grammar

in schemes known to differ from traditional grammatical models. [24] As described in the introduction, the corpus and language model used for this analysis have been developed and are presently used by Oxford University Press, one of the most recognized authorities on the English language and is thereby assumed to map well to well-recognized models of English grammar.

Fig 2 shows the relationships found to be most prevalent, grouped according to traditional grammatical classification, and listed according to prevalence. These findings are elucidated below.

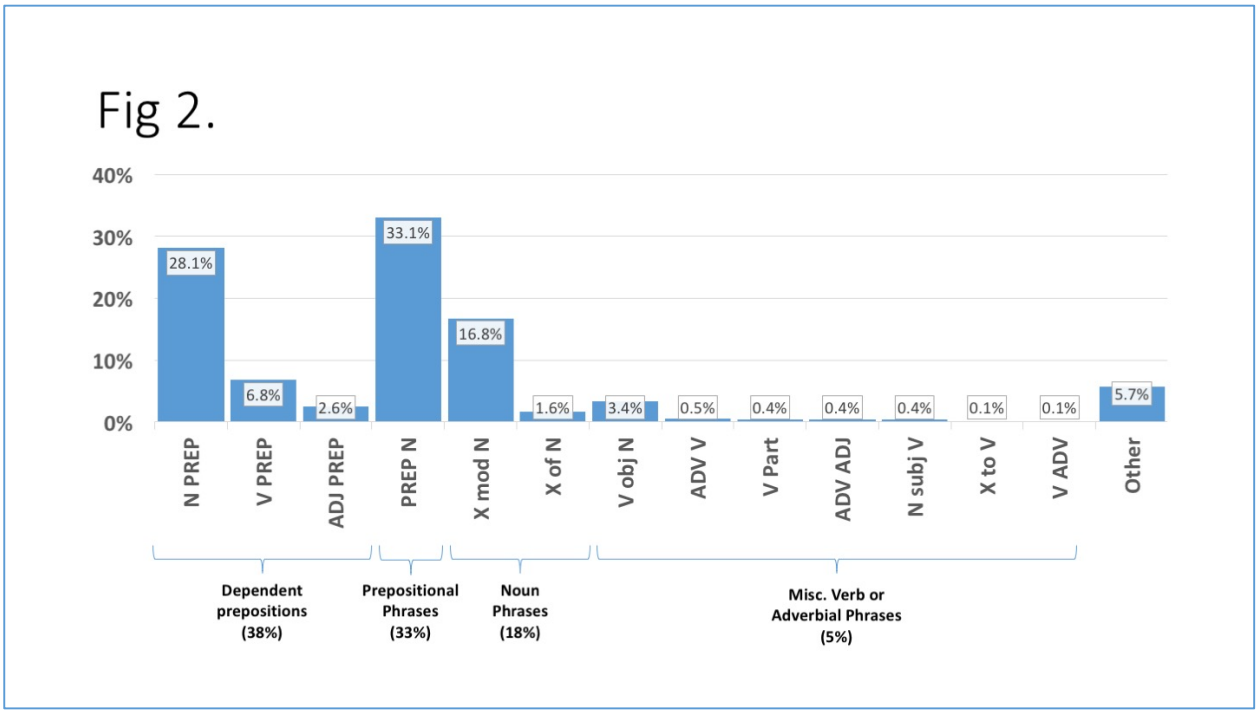

Fig 2. Grammar relationships, grouped according to POS. Y-axis shows prevalence as a percentage of the aggregate sum (n=9561). X-axis lists the various grammatical relationships included for discussion, grouped according to high-level grammatical part-of-speech.

Dependent prepositions (38%)

Dependent prepositions come in three varieties, dependent nouns, dependent verbs, and dependent adjectives and are represented in our data by the "N Prep", "V Prep", and "Adj Prep" relationships.

Together, this grammatical category was found to have a combined prevalence of 38%. Dependent prepositions are word forms in which the meaning of the respective POS depends entirely upon the attached preposition, and strongly reflect the ancient roots of medical English. For nNS, these forms are particularly difficult to master due to the incorrect assumption that the meaning of the dependent preposition directly reflects the combined meaning of the constituent POS and prepositions. Indeed, in languages, such as German, where separable POS are more commonly recognized within the pedagogy, dependent constructions such as these are learned as wholly independent terms that are separable according to strict grammatical conventions. The high prevalence of dependent prepositions within the body of medical English strongly implies the need for a similar approach to be adopted by learners and instructors charged with their learning.

Prepositional phrases (33%)

The high-incidence prepositional phrases found within medical English generally represent subordinate clauses that refer to additional information commonly required within medical discourse. Unlike dependent prepositions, prepositional phrases can and should be understood as separable units, with a combined meaning that can be logically inferred from the constituent parts. That said, the high prevalence of such phrases requires that learners wishing to function competently at this level be familiar with and comfortable using these patterns. This is the hallmark of high-level fluency, which has been shown to directly impact perceived credibility. [25] In fields such as medicine, where credibility is essential to the life and well-being of others, instructors and learners alike have a duty to ensure adequate familiarity with and competence using these terms.

Noun phrases (18%)

Nouns, which are commonly assumed to be the most essential component of medical English education, were found to be overall less relevant with respect to improving learners' overall medical English capacity. Indeed, aside from the overall lower prevalence within medical English, correct usage and interpretation requires a less intimate knowledge of context or collocational usage. In addition, given the overall lower likelihood that any given noun phrase will be encountered in a given text (see Table 2, term frequency), learners and instructors may be better served allocating efforts to mastery of the other constructs previously discussed in this section. Indeed, as the present authors have observed with their own students, high-aptitude learners (such as doctors and medical students) can easily assimilate new vocabulary encountered in context, especially with the help of electronic dictionaries and instantaneous, online information retrieval services such as Google. This observation is corroborated by previous studies supporting the effectiveness of this approach. [26]

Miscellaneous verb or adverbial phrases (5%)

Verbs play a central role in the understanding and usage of any language, so much so in fact that the verbs most commonly found in medical English are also relatively common overall. This is reflected by the comparably lower incidence of verbs and verb phrases in the present study, which systematically identified and extracted only terms found to be relatively more prevalent within medical English, as opposed to general usage. Consequently, this category is dominated by "V obj N" compound verb constructions that reflect specialized usage of otherwise common POS. Examples include "have … effect", "play … role", "treat … patient", "provide … information", "reduce … risk", and "make … decision". Table 3 provides examples of typical usage. As is clearly

evident, these terms behave as and may be more appropriately understood to be separable verbs similar, in principle, to those found commonly in languages such as German. And as previously discussed, a pedagogy that treats such terms accordingly may ultimately be most effective for ensuring competence at this level. As shown, these terms are composed of relatively common components, making learner error highly likely if no intervention is made. Therefore, given that this list is comprised of only 281 items, we feel it would be remiss to omit these from instruction.

Table 3. Example usage for "V obj N".

V obj N	Example
have ... effect	Vitamin E supplementation [[had no apparent effect]] on basal endothelial
	When the mutation is homozygous it [[has a much greater effect]] , and embryos
	Thus , the effect of extubation in this subset of patients [[had an appreciable clinical effect]] .
play ... role	Upbringing [[plays an important role]] .
	Zinc [[plays a vital role]] in connective
	availability of nonfusion technologies will likely [[play a significant role]] in changing the
treat ... patient	We [[treated 114 patients]] with RIF / PZA
	When glaucoma is confirmed , the [[patient is medically treated]] and is reexamined
	that you can never [[treat a patient]] with a borderline
provide ... information	comprehension is primarily dependent on the [[information provided]] explicitly within the
	We will continue to follow events and [[provide information]] as it comes around
reduce ... risk	You can [[reduce the risk]] of stomach upset
	You can [[reduce this risk]] by managing your
	While a preventive mastectomy [[reduces your risk]] of breast cancer
make ... decision	Women considering taking HRT should [[make that decision]] with their clinician
	to help both of you communicate and [[make important decisions]] , it can be
	While many physicians want to [[make decisions]] guided by the best

Table 3. Example usage for "V obj N". Examples above represent a sample of the data set corresponding to the grammatical relationship, "V obj N". For this group, passive constructions are shown to demonstrate collocational behavior identical to their active construction counterparts.

In addition to the "V obj N" items described above, this research identified six other classes of verb-related relationships. Unlike "V obj N", most are simple collocations that happen to be more prevalent in medical discourse. Due to their low incidence, these groups are excluded from detailed discussion; however, the preceding conclusions generally apply. With only 137 items, learners are urged to become at least minimally familiar. Supplement 1 provides a comprehensive overview.

Practical implications

In addition to identifying and characterizing the grammatical relationships most prevalent in medicine, this research also identifies the specific terminology without which medical English cannot be appropriately understood or used. Most of these terms have usage and meaning that, within the medical context, varies substantially from what would be the otherwise typical interpretation. Without a clear understanding of contextual variation or what constitutes typical usage, nNS may be more prone to errors in communication or interpretation. [27] Furthermore, as has been demonstrated by a substantial body of previous research, without the ability to appropriately recognize and use idiomatic expressions, nNS cannot communicate fluently or function appropriately at the advanced level of proficiency required in the professional setting. [28]

Accordingly, Fig 3 incorporates all data previously discussed and maps it according to average frequency per term, average prevalence per term, and overall term count. This visualization suggests a framework for prioritizing medical English learning, in which the highest-yield learning strategy is shown to be one which focus primarily on dependent preposition patterns (i.e., "N Prep", "Prep N", "V Prep", and "Adj Prep") and usage of prepositional phrases ("Prep N"). In addition, this visualization poignantly highlights the finding that noun phrases ("X mod N"), while highly important as a whole, are individually far less likely to be encountered in any given context.

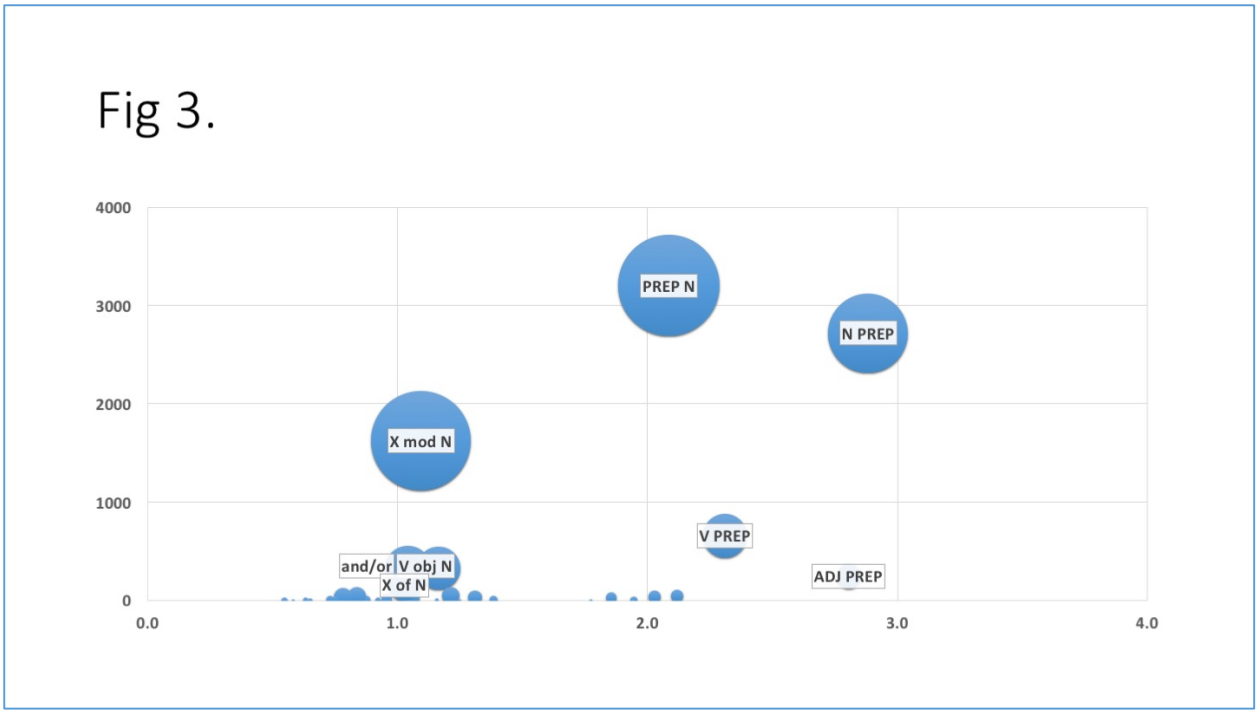

Fig 3. Overview of key grammar relationships. X-axis maps average term frequency per grammar relationship (importance), while Y-axis maps prevalence (term count * frequency). Bubble size represents term count (challenge). Only categories with term count > 100 are labelled.

As previously discussed, noun phrases ("X mod N", "V obj N", and "X of N"), the more typical focus of medical English learning material, are comprised of terms that are 2-3 times less likely to be encountered in any given text. These results shed light on the seemingly paradoxical situation in which vocabulary building, while acknowledged to be essential, is generally not regarded as the most productive use of learning time. [29] Indeed, it is well known that beyond article usage, the most common error for nNS relates to the usage of prepositions and prepositional phrases. [30] And interestingly, the present research found these to be the most prevalent within medical English.

Limitations

This research pre-supposes that the corpus used was sufficiently representative for the purpose of population-level inference. As of the time of this writing, this cannot be confirmed. However, as previously discussed, the OEC is considered to be the most comprehensive and highest quality English-language corpus presently available. And it has been employed by numerous authorities for research into current English usage. Thus the authors contend that the results demonstrated, if not statistically robust, are nevertheless sufficiently accurate and precise with respect to our stated aims. Moreover, only aggregated data was subject to analysis; no assertions were made regarding the importance of individual terms or phrases. Consequently, we expect these present findings to be highly reproducible and unlikely to vary significantly with the introduction of a new or updated corpus of comparable or superior quality.

Conclusion

In the present study, computational linguistics methods have been used to identify the prevalence of key terms and phrases essential to the understanding of medical English. By systematically identifying such key terms and phrases, we were able to more precisely characterize not only the words out of which medical English is comprised, but also the logic and grammar most associated with this highly specialized field. The data presented in this study has strong implications regarding how to most efficiently improve the communication competence of IMGs, as well as students and doctors intending to work in countries where English is not the first language. By developing targeted teaching sessions focusing on preposition-dependent terms as opposed to crude medical vocabulary, these findings can form the basis for a prospective case-control study to analyze the effect of these two different strategies on future doctor-patient and doctor-doctor interactions.

295

Footnotes

Acknowledgements:

We thank Prof Esteban Gabazza and Dr. Corina Gabazza for their kind advice and encouragement throughout the research and drafting process. We also thank the Mie University Center for Medical and Nursing Education for approving and supporting this research. Special thanks go to the team at Oxford University Press for providing access to their data and for supporting this research at several key steps, as well as the team at SketchEngine for their kind and patient support throughout. We would also like express our deepest appreciation to Prof Dr Daisy Rotzoll for providing strategic insights regarding the implications and practical application of this research, as well as for providing suggestions for several key revisions to the final manuscript.

Contributors:

AG designed the study, conducted relevant data preparation and analyses, created all figures and tables, and drafted and edited the paper.

Competing Interests:

All authors have completed the ICMJE uniform disclosure form and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Transparency Declaration:

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

References

1 OECD W. International Migration of Health Workers Improving International Co-Operation To Address The Global Health Workforce Crisis Brief Paris. 2010.

2 OECD. International migration of doctors. In: *Health at a Glance*. Organisation for Economic Co-operation and Development 2015. 86–7.http://www.oecd-ilibrary.org/content/chapter/health_glance-2015-24-en (accessed 29 Dec2015).

3 Lineberry M, Osta A, Barnes M, *et al*. Educational interventions for international medical graduates: a review and agenda. *Med Educ* 2015;**49**:863–79. doi:10.1111/medu.12766

4 Fiscella K, Roman-Diaz M, Lue BH, *et al*. 'Being a foreigner, I may be punished if I make a small mistake': assessing transcultural experiences in caring for patients. *Fam Pract* 1997;**14**:112–6. doi:10.1093/fampra/14.2.112

5 Searight HR, Gafford J. Behavioral science education and the international medical graduate. *Acad Med* 2006;**81**:164–70.

6 Dorgan KA, Lang F, Floyd M, *et al*. International Medical Graduate–Patient Communication: A Qualitative Analysis of Perceived Barriers: *Acad Med* 2009;**84**:1567–75. doi:10.1097/ACM.0b013e3181baf5b1

7 Wilner LK, Feinstein-Whittaker M. Improving Communication Skills in Health Care. *SIG 14 Perspect Commun Disord Sci Cult Linguist Diverse CLD Popul* 2013;**20**:109–17.

8 Staples S. Examining the linguistic needs of internationally educated nurses: A corpus-based study of lexico-grammatical features in nurse–patient interactions. *Engl Specif Purp* 2015;**37**:122–36. doi:10.1016/j.esp.2014.09.002

9 Man JP, Weinkauf JG, Tsang M, *et al*. Why do some countries publish more than others? An international comparison of research funding, English proficiency and publication output in highly ranked general medical journals. *Eur J Epidemiol* 2004;**19**:811–7.

10 Charlton BG. How can the English-language scientific literature be made more accessible to non-native speakers?: Journals should allow greater use of referenced direct quotations in 'component-oriented' scientific writing. *Med Hypotheses* 2007;**69**:1163–4. doi:10.1016/j.mehy.2007.07.007

11 Vasconcelos SMR, Sorenson MM, Leta J. Scientist-friendly policies for non-native English-speaking authors: timely and welcome. *Braz J Med Biol Res* 2007;**40**:743–7. doi:10.1590/S0100-879X2007000600001

12 Stretton S, Bramich NJ, Keys JR, *et al*. Publication misconduct and plagiarism retractions: a systematic, retrospective study. *Curr Med Res Opin* 2012;**28**:1575–83. doi:10.1185/03007995.2012.728131

13 Ho RC-M, Mak K-K, Tao R, *et al*. Views on the peer review system of biomedical journals: an online survey of academics from high-ranking universities. *BMC Med Res Methodol* 2013;**13**:74. doi:10.1186/1471-2288-13-74

14 Alcaraz Ariza MÁ, others. The English of the health sciences: a note on foreign borrowings. Published Online First: 2012.http://rua.ua.es/dspace/handle/10045/35748 (accessed 20 Nov2015).

15 Gledhill CJ. *Collocations in science writing*. Gunter Narr Verlag 2000.

16 The Oxford English Corpus - Oxford Dictionaries. http://www.oxforddictionaries.com/words/the-oxford-english-corpus (accessed 11 Dec2015).

17 Comparison of the COCA and OEC. http://corpus.byu.edu/coca/help/compare_oec.asp (accessed 11 Dec2015).

18 Sketch Engine | Home page. https://www.sketchengine.co.uk/ (accessed 11 Dec2015).

19 Kilgarriff A. Terminology finding, parallel corpora and bilingual word sketches in the Sketch Engine. In: *Proc ASLIB 35th Translating and the Computer Conference, London*. 2013. http://www.mt-archive.info/10/Aslib-2013-Kilgarriff.pdf (accessed 11 Dec2015).

20 Taylor A, Marcus M, Santorini B. The Penn treebank: an overview. In: *Treebanks*. Springer 2003. 5–22.http://link.springer.com/chapter/10.1007/978-94-010-0201-1_1 (accessed 17 Dec2015).

21 Dunning T. Accurate Methods for the Statistics of Surprise and Coincidence. *Comput Linguist* 1993;**19**:61–74.

22 Kilgarriff A. Getting to Know Your Corpus. In: Sojka P, Horák A, Kopeček I, *et al*., eds. *Text, Speech and Dialogue*. Springer Berlin Heidelberg 2012. 3–15.http://link.springer.com/chapter/10.1007/978-3-642-32790-2_1 (accessed 11 Dec2015).

23 Kilgarriff A, Husák M, McAdam K, *et al*. GDEX: Automatically finding good dictionary examples in a corpus. In: *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*. 2008. 425–32.http://dialnet.unirioja.es/servlet/articulo?codigo=5040252 (accessed 11 Dec2015).

24 Kehler A, Kehler A. *Coherence, reference, and the theory of grammar*. CSLI publications Stanford, CA 2002.

25 Lev-Ari S, Keysar B. Why don't we believe non-native speakers? The influence of accent on credibility. *J Exp Soc Psychol* 2010;**46**:1093–6. doi:10.1016/j.jesp.2010.05.025

26 Hulstijn JH. Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. *Vocab Appl Linguist* 1992;:113–25.

27 Lev-Ari S, Keysar B. Less-Detailed Representation of Non-Native Language: Why Non-Native Speakers' Stories Seem More Vague. *Discourse Process* 2012;**49**:523–38. doi:10.1080/0163853X.2012.698493

28 Laufer B, Waldman T. Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Lang Learn* 2011;**61**:647–72. doi:10.1111/j.1467-9922.2010.00621.x

29 Townsend D, Kiernan D. Selecting Academic Vocabulary Words Worth Learning. *Read Teach* 2015;**69**:113–8. doi:10.1002/trtr.1374

30 Han N-R, Chodorow M, Leacock C. Detecting errors in English article usage by non-native speakers. *Nat Lang Eng* 2006;**12**:115–29. doi:10.1017/S1351324906004190

Appendix

Supplemental Table 1. Overview of miscellaneous verbs and adverbial phrases.

rank	ADV V	N subj V	ADV ADJ	V Part	X to V
1	also ... have	patient ... have	statistically ... significant	follow ... up	likely ... to ... have
2	also ... find	study ... show	significantly ... high	carry ... out	use ... to ... treat
3	commonly ... use	child ... have	significantly ... low	go ... on	find ... to ... have
4	also ... show	patient ... receive	significantly ... different	point ... out	use ... to ... assess
5	sexually ... transmit	study ... find	so ... many	find ... out	have ... to ... do
6	also ... include	study ... suggest	very ... important	set ... up	use ... to ... determine
7	significantly ... reduce	% ... have	very ... low	rule ... out	appear ... to ... have
8	widely ... use	woman ... have	as ... opposed	make ... up	want ... to ... know
9	significantly ... increase	people ... have	very ... high	pick ... up	likely ... to ... develop
10	also ... provide	researcher ... find	too ... much	turn ... out	need ... to ... treat
11	also ... know	research ... show	critically ... ill	come ... up	show ... to ... have
12	previously ... report	study ... demonstrate	significantly ... great	end ... up	intention ... to ... treat
13	well ... know	patient ... undergo	very ... good	give ... up	…
14	randomly ... assign	patient ... take	as ... high	take ... up	likely ... to ... report
15	also ... use	study ... report	as ... effective	grow ... up	show ... to ... reduce
16	often ... have	patient ... experience	much ... high	break ... down	
17	also ... report	data ... suggest	very ... different	come ... in	
18	still ... have	study ... use	very ... small	work ... out	
19	also ... help	study ... examine	relatively ... small	build ... up	
20	also ... suggest	study ... indicate	very ... difficult	come ... out	
21	well ... understand	% ... report	as ... possible		
22	often ... use	patient ... need	as ... likely		
23	previously ... describe	patient ... develop	very ... similar		
24	…	finding ... suggest	too ... many		
25	also ... increase	patient ... die	relatively ... low		
26	et_al ... .find	patient ... report	commercially ... available		
27	well ... tolerate	patient ... require	minimally ... invasive		
28	randomly ... select	symptom ... include	significantly ... related		
29	also ... need	study ... compare			
30	also ... call	patient ... present			
31	also ... occur	infection ... cause			
32	et_al ... .report	patient ... use			
33	already ... have				
34	also ... note				
35	now ... have				
36	significantly ... decrease				
37	also ... associate				
38	usually ... occur				
39	strongly ... associate				
40	poorly ... differentiate				
41	newly ... diagnose				
42	significantly ... correlate				