# Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis

Stephen T Wu,[1] Hongfang Liu,[1] Dingcheng Li,[1] Cui Tao,[1] Mark A Musen,[2] Christopher G Chute,[1] Nigam H Shah[2]

[1]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA
[2]Stanford Center for Biomedical Informatics Research, Stanford, CA, USA

**Correspondence to**
Dr Stephen T Wu, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA;
wu.stephen@mayo.edu

Received 2 December 2011
Accepted 12 March 2012
Published Online First
4 April 2012

## ABSTRACT

**Objective** To characterise empirical instances of Unified Medical Language System (UMLS) Metathesaurus term strings in a large clinical corpus, and to illustrate what types of term characteristics are generalisable across data sources.

**Design** Based on the occurrences of UMLS terms in a 51 million document corpus of Mayo Clinic clinical notes, this study computes statistics about the terms' string attributes, source terminologies, semantic types and syntactic categories. Term occurrences in 2010 i2b2/VA text were also mapped; eight example filters were designed from the Mayo-based statistics and applied to i2b2/VA data.

**Results** For the corpus analysis, negligible numbers of mapped terms in the Mayo corpus had over six words or 55 characters. Of source terminologies in the UMLS, the Consumer Health Vocabulary and Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) had the best coverage in Mayo clinical notes at 106 426 and 94 788 unique terms, respectively. Of 15 semantic groups in the UMLS, seven groups accounted for 92.08% of term occurrences in Mayo data. Syntactically, over 90% of matched terms were in noun phrases. For the cross-institutional analysis, using five example filters on i2b2/VA data reduces the actual lexicon to 19.13% of the size of the UMLS and only sees a 2% reduction in matched terms.

**Conclusion** The corpus statistics presented here are instructive for building lexicons from the UMLS. Features intrinsic to Metathesaurus terms (well formedness, length and language) generalise easily across clinical institutions, but term frequencies should be adapted with caution. The semantic groups of mapped terms may differ slightly from institution to institution, but they differ greatly when moving to the biomedical literature domain.

## BACKGROUND AND SIGNIFICANCE

Natural language processing (NLP) is crucial to clinical informatics because the summative information that is stored in millions of clinical notes is too massive to be processed by a human. But automatic methods of processing clinical text have their own challenges, such as the extensive use of specialised medical terms. The Unified Medical Language System (UMLS) Metathesaurus[1] has over 8 million strings that an NLP system might consider relevant in clinical text. It is thus common practice for NLP systems[1 2] to filter the desired terms by criteria such as lexical redundancy and term ambiguity[2] or semantic type.[3] Such filters, while reasonable, are uninformed by how the terms behave in clinical text.

The long-term goal of this work is to produce an agile information extraction user interface that allows users to specify terms, concepts and logic relevant to their own problem settings, based on criteria such as frequency, source terminology, syntax and semantic type. To that end, our objective here is twofold: first, to analyse empirical instances of UMLS term strings in a large clinical corpus; and second, to illustrate what types of term characteristics are generalisable across data sources. The resulting statistics and principles may then be used in user-directed filtering of lexicons (eg, using Lexicon Builder[4]) for practical clinical NLP systems. This may also improve system efficiency—the full Metathesaurus (prohibitively, for some users) requires several gigabytes of memory to serve as a lexicon for many algorithms.

This paper therefore explores the characteristics of Metathesaurus term matches in clinical text along dimensions such as term length, term frequency, source terminology, syntactic category and semantic group. The data source used is a corpus of over 51 million patient notes gathered over a 10-year period at the Mayo Clinic. A variant of the standard Aho-Corasick string matching algorithm[5 6] is run on the data to find term matches, and these data are paired against existing information from Mayo's enterprise NLP system, Clinical Notes Indexing (CNI),[7] a precursor to Mayo's open-source NLP system, cTAKES.[3] The paper also examines the transferability of corpus statistics by applying a set of Mayo-based filtering parameters to the i2b2/VA NLP Challenge corpus.[8] This cross-institutional test provides some insight on which statistical metrics are mainly beneficial within one setting and which are broadly applicable.

After a brief discussion on related work, the remainder of this article introduces the data and methods for empirical term matching in clinical corpora, analyses the Mayo Clinic corpus of clinical notes, applies and analyses a practical set of filters and draws a few conclusions for NLP tasks.

## RELATED WORK

The UMLS Metathesaurus[1] is constantly growing as its source terminologies grow; its 2011AA release contains 155 sources with 8 335 125 different strings for terms in 21 languages, and 2 404 937 different concept unique identifiers. As a thesaurus, the Metathesaurus is designed to match identical concepts from different source terminologies, and it has thus been used frequently as a normalisation target for NLP methods.[3 7 9–11] Our previous work has analysed the large-scale distribution of UMLS clinical concepts.[12]

The Metathesaurus has also been commonly used as a lexicon[2] to supply term strings that might be identified in clinical text, which is slightly different than the concept-oriented focus for which it was designed. This incongruency has been addressed to some degree in MetaMap, an NLP system from the National Library of Medicine, which allows some configurable filtering of the lexicon.[2 13] This filtering is helpful, but lacks the ability to provide a user with in-domain, empirically based recommendations. With the rise in computational power and the increasing availability of biomedical ontologies, we believe that a corpus-driven approach[14] is feasible for principled lexicon filtering.

Constructing practical string-oriented lexicons through filtering has been attempted via statistical models and via rule-based systems. Statistical models typically identify a number of properties that allow prediction of the likelihood of a given string being found or not found in a corpus.[15] An excellent recent rule-based study by Hettne *et al*[16] recommends applying five rewrite rules (of nine studied) and seven suppression rules (of eight studied) to the UMLS before it is used for biomedical term identification in MEDLINE.[16] Our work complements these attempts by highlighting the large-scale effects of the lexicon-building technique of term suppression.

In the biomedical literature domain, the efforts at lexicon creation are quite advanced; for example, the BioLexicon gathers terms from existing data resources into a single, unified repository, and augments them with new term variants extracted from biomedical literature.[17] Efforts by Baral *et al* provide an online dictionary of diseases and drugs based on frequency analysis in Medline (http://bioai4core.fulton.asu.edu/snpshot/download.html). Our work in analysing a large-scale clinical corpus provides a principled foundation for creating such resources in the clinical domain.

Other corpus studies have been conducted which analyse variability in subdomains,[18] sections of a document,[19] large-scale semantic characteristics of biomedical literature abstracts,[20] and longitudinal semantic shift.[21] Our previous work also includes comparisons between concepts in the clinical and biomedical domains.[12] Here, we undertake the first known enterprise-scale exploration of clinical text that centres on term strings actually present in the text.

## DATA AND METHODS

### Data sources

The data source for the corpus analysis of clinical text was Mayo Clinic clinical notes between 1 January 2001 and 31 December 2010, retrieved from the Mayo's Enterprise Data Trust (EDT).[22] The EDT stores structured data, unstructured text and CNI-produced annotations[7] from a comprehensive snapshot of Mayo Clinic's service areas, excluding only microbiology, radiology, ophthamology and surgical reports. Additionally, each possible note type at Mayo was represented: clinical note, hospital summary, post-procedure note, procedure note, progress note, tertiary trauma and transfer note.

For the evaluation of a sample filter, the i2b2/VA 2010 NLP Challenge data[8] were used. This corpus contained a total of 871 manually annotated, de-identified reports from Partners Healthcare, Beth Israel Deaconess Medical Center and the University of Pittsburgh Medical Center. The majority of notes were discharge summaries, but the University of Pittsburgh Medical Center also contributed progress reports.

### String matching algorithm

Our string matching procedure implemented a modified Aho-Corasick algorithm.[5] This algorithm takes a dictionary and constructs a finite state machine with efficient transitions between alphabet string states for failed matches. Our modification uses normalised words as the alphabet, but we store the original strings for each match and report results on exact matches.

We used the UMLS Metathesaurus as a lexicon. Due to computational constraints we filtered out entries with 10 or more words and those that were not between 3 and 100 characters. Because the algorithm used the UMLS Metathesaurus there were concept unique identifiers available for each string match. We used this normalised representation to find type unique identifiers and characterise the semantic types of the strings.

### Data collection and preparation

For corpus analysis, we retrieved text documents from the EDT repository, with 51 945 627 documents represented from 2000 to 2010. The dictionary lookup procedure described above found any UMLS terms in the text documents. For analysis by syntactic category, we retrieved CNI-produced syntactic chunks[7] for the same set of documents, and the dictionary lookup procedure was applied to the text of these chunks. This yielded the syntactic category for the majority of term occurrences in the text.

For the last step of examining the cross-institutional transferability of statistics, we used the 2010 i2b2/VA NLP Challenge data without modification. As above, the dictionary lookup procedure mapped UMLS terms in the i2b2/VA data.

## RESULTS AND ANALYSIS

### Corpus analysis

#### Aggregate characteristics

In the corpus of 51 945 627 clinical documents, there are a total of 2 319 010 575 case-insensitive exact term matches, drawing from 296 167 unique terms. This amounts to 44.64 matches per document on average and only utilises 3.56% of the available case-insensitive terms in the UMLS. It is thus clear that we do not need to search the full Metathesaurus in the course of a concept mapping procedure.

However, we should not overestimate how much the terminologies may be filtered, as the dictionary lookup algorithm used was fairly unsophisticated. In fact, it is unlikely that there are so few terms per document in clinical text. Xu *et al* report 19 million Medline abstracts to have 530.45 matched terms per document using 13% of the unique strings in the UMLS.[20] This difference is particularly stark in light of the fact that the clinical documents have, on average, three times as many characters (about 2500) as biomedical abstracts.

The larger number of biomedical matches is likely indicative of the fact that the biomedical text covers a broader range of topics than clinical text. It is also difficult for exact dictionary matches to fully capture the range of synonymous expressions, abbreviations and misspellings that are found in clinical text. For example, the strings 'dispo' (abbreviation for the disposition of a patient) and '00Cardiac implant' (tokenisation problems) both occur in the Mayo corpus but are not identifiable.

All of these factors point to a large difference between the clinical and biomedical domains, and also to the need for a clinical data-specific study such as this one.

#### Word and character statistics

As previously mentioned, the UMLS Metathesaurus was designed as a controlled thesaurus not a lexicon. It therefore

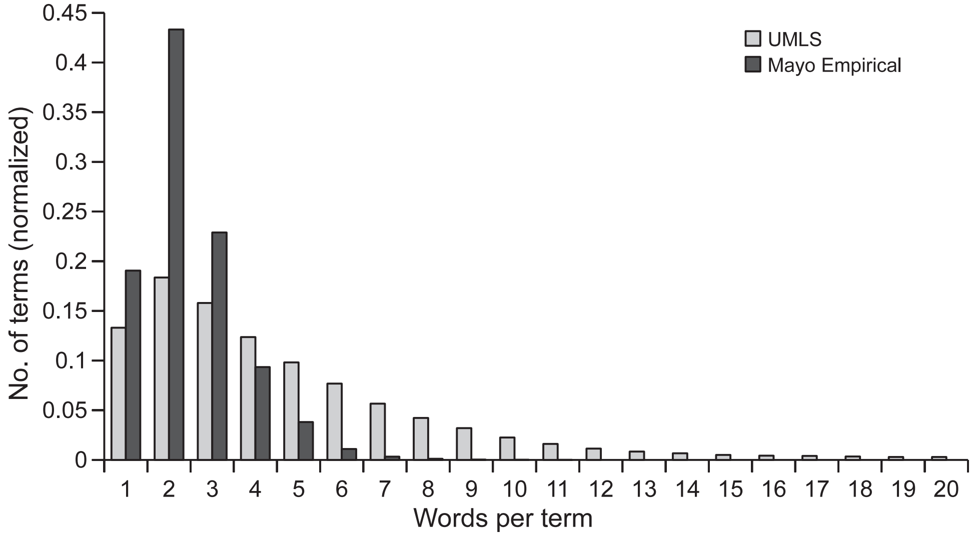

**Figure 1** The number of words in a term versus relative frequency of Unified Medical Language System (UMLS) terms with that number of words.

contains concepts that include an excessive number of words or characters and are not of use to NLP techniques. Figure 1 shows histograms for the number of words in the UMLS and in the subset that is empirically found in Mayo Clinic data.

It should be clear that the mappable dictionary terms from the UMLS are shorter on average than the full set of UMLS terms. Subsetting to these 296 167 terms reduces the average characters per term from 37.27 to 17.83 and average words per term from 4.80 to 2.41, similar to the characteristics reported in the biomedical domain. The same is seen to be true when examining the number of characters in UMLS terms, as in figure 2.

These findings suggest that filtering out high word counts or character counts may be a safe way to remove unnecessary terms from a lexicon.

### Term frequency and TF−IDF

To understand what types of UMLS strings are found in clinical text, we now consider some traditional metrics for the importance of a term. Figure 3 shows the distribution of the top 5000 term frequencies in each domain.

We have scaled the y-axis for biomedical term frequencies to be comparable with the clinical domain. The x-axis is ordered by term frequency (tf) ranking, where the top strings are seen in table 1A,B. We can see that few terms are used frequently (the left portion of figure 3) and many terms are used infrequently (the bottom/right portion), and this characteristic is consistent across both domains. This is reminiscent of Zipf's Law, which describes the empirical frequency distribution of words in general language as having a large peak and a heavy, one-sided tail. By the technical log−log plot definition of a Zipfian distribution, we would see that this is near-Zipfian but the tail is not as heavy.

From table 1 it is evident that in both domains, the most frequent terms are general rather than specific, and reflect the domains from which they arise. In 51 million documents, 7.7% of terms only occurred once; the 0%, 25%, 50%, 75% and 100% quantiles are at 1, 3, 18, 85 and 38 434 437 occurrences, respectively.

We additionally obtained the tf−idf weight of each term for the clinical corpus as in table 2. Tf−idf weights are defined by $tf - df = tf \cdot \log(|N|/df)$, where n is the number of documents in the corpus and df is the number of documents a term occurs in. They are commonly used in information retrieval to measure the importance of terms, with the intuition that terms that occur often in every document are less distinctive than those that occur often in a few documents. Note that the top terms are very similar to the term frequency-ranked versions.

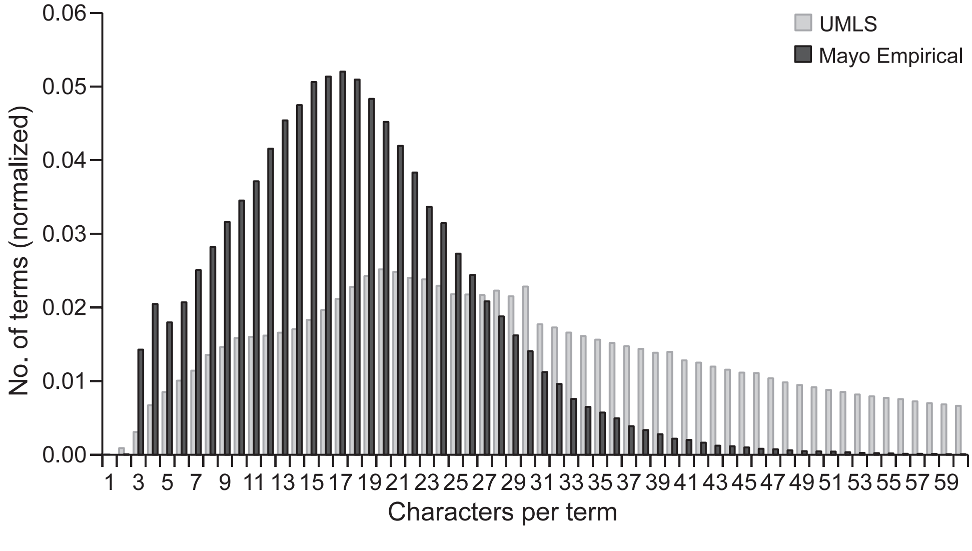

**Figure 2** The number of characters in a term versus how many Unified Medical Language System (UMLS) terms had that number of characters.

**Figure 3** Distribution of the most frequent terms in clinical versus biomedical data.

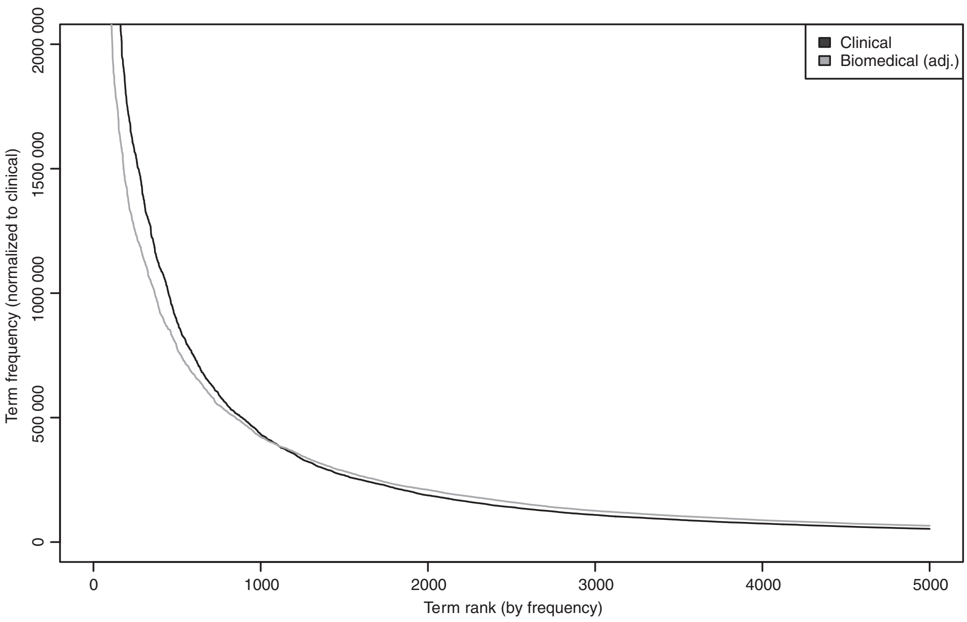

Figure 4 visualises this comparison by showing the tf rank (x-axis) with the tf−idf values (y-axis)—they are still highly consistent. From here, we see that traditional information retrieval metrics such as tf−idf may be somewhat limited in their ability to discover truly valuable, discriminative words in the clinical domain.

This ineffectiveness of inverse document frequency is likely due to the fact that the clinical domain is highly specialised by note type and subdomain. The term 'patient' is discriminative in some respects: it can be easily found in progress notes and discharge summaries, but is much less likely to be found in notes like pathology or radiology reports.

### Source terminology

Here, we compare the number of strings per terminology in the raw UMLS (table 3A) with the most commonly used terminologies (by number of terms represented) in the clinical and biomedical domains (table 3B,C).

These tables show which terminologies are best for each domain, ranked by the number of unique case-insensitive terms used. Tables 3B and 3C also include what percentage of the terms in the full terminology are used. Interestingly, the new Consumer Health Vocabulary contains only 148 383 terms but accomplishes excellent coverage of terms in both domains because it was designed for natural language contexts. The Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) is the largest source ontology in the UMLS and was developed specifically as a clinical resource. As such, it is one of the most important terminologies in the clinical domain. Similarly, Medical Subject Headings (MSH) was developed specifically for indexing biomedical literature and therefore captures the most terms from biomedical abstracts.

The percentage usage of each of these ontologies is lower in the clinical domain than in the biomedical domain, again likely due to applying an exact case-insensitive string match to highly varied clinical notes. Low usage rates in the clinical domain, for example, SNOMED-CT, also indicate that the resource may best contribute to a lexicon after some filtering along other dimensions.

### Semantic groups

As mentioned above, the frequent words in the clinical domain differ from those in the biomedical domain. This is most easily seen in figure 5A,B.

The percentages of matched strings are compared by semantic group and they differ greatly. Here, we follow Bodenreider and McCray's 15 semantic groups[23] of semantic types (UMLS Type Unique Identifiers) figure 6.

These plots display predictable domain differences in semantic type distribution of terms. Clinical data focus on disorders, anatomy, medications and procedures. cTAKES and CNI are examples of intentional semantic type-based filtering for clinically relevant types, in which five semantic groups are kept,

**Table 1** Top terms in clinical text (Mayo corpus) and biomedical text (Medline 2011), by term frequency

| (A) Clinical text | | (B) Biomedical text | |
|---|---|---|---|
| **Term** | **Frequency** | **Term** | **Frequency** |
| Patient | 38 434 437 | Patients | 10 393 786 |
| Not | 18 601 179 | Cells | 4 855 359 |
| History | 16 650 248 | Treatment | 4 103 013 |
| Pain | 15 125 464 | Study | 4 032 105 |
| Time | 14 667 600 | Results | 3 498 940 |
| Normal | 13 656 279 | Cell | 3 082 455 |
| Right | 13 181 157 | Using | 2 840 963 |
| Left | 13 170 124 | Effect | 2 754 055 |
| Daily | 10 923 371 | Activity | 2 610 750 |
| Well | 9 534 581 | Protein | 2 332 732 |

**Table 2** Top terms in clinical text by tf−idf weight

| Term | Frequency | Document frequency | tf−idf |
|---|---|---|---|
| Patient | 38 434 437 | 12 163 186 | 5.5E+07 |
| Not | 18 601 179 | 6 921 338 | 3.7E+07 |
| Pain | 15 125 464 | 4 883 178 | 3.5E+07 |
| History | 16 650 248 | 7 375 392 | 3.2E+07 |
| Normal | 13 656 279 | 5 265 335 | 3.1E+07 |
| Daily | 10 923 371 | 2 984 235 | 3.1E+07 |
| Right | 13 181 157 | 5 351 140 | 3.0E+07 |
| Left | 13 170 124 | 5 388 304 | 3.0E+07 |
| Time | 14 667 600 | 7 177 814 | 2.9E+07 |
| Day | 8 288 472 | 3 358 834 | 2.3E+07 |

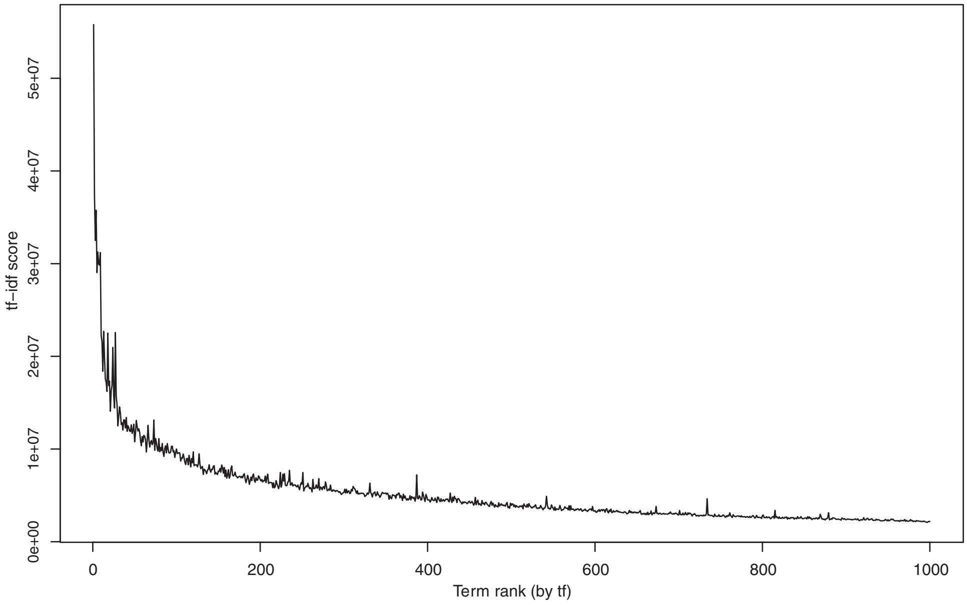

**Figure 4** Tf−idf values of the most frequent terms in clinical data.

accomplishing 59.60% coverage of occurrences and 82.74% coverage of unique strings.

Note that the difference between the clinical and biomedical domains is very significant. Type filters designed for one domain should not be applied to another, though some semantic groups are relatively infrequent to both domains.

### Syntactic categories

Across the Mayo clinical notes in this study, we found that Across the Mayo clinical notes in this study, 90.18% of clinical term mentions were found in noun phrase (NP) chunks; Xu *et al* found similar NP-dominance characteristics in biomedical data. Figure 6 stratifies the clinical NP-dominance characteristics by semantic group. While filtering out non-NP constructions is commonplace in many clinical NLP systems, it should be done with caution in for semantic groups like “Procedures” or “Activities & Behaviors”.

It should be noted that this depends on a sound chunking procedure, and there were some limitations to the accuracy of the IBM shallow parser in CNI: there were terms that resided in incorrect chunks and those that were not in any chunk. However, as string-matched terms occur across the whole distribution of the text, this noise is overcome on average.

## Cross-institutional analysis

Based on the corpus analysis on Mayo data above, we defined an example configuration of filters for use-case agnostic information extraction in clinical notes, and applied these candidate filters to string-matched i2b2/VA data to examine their trans-institutional applicability.

### A Mayo-based filtering configuration

We implemented eight lexicon filters:

1. *Special characters.* The UMLS contains fine-grained semantic distinctions that are indicated with punctuation, for example, ‘[D] Respiratory insufficiency (finding)’ versus ‘Respiratory insufficiency, NOS.’ This UMLS-intrinsic filter removes a term from the lexicon if and only if it begins with ‘[’ ends with‘)’ or contains a comma.[20]
2. *Maximum number of words.* Given the histogram in figure 1, fewer than 1000 terms have seven words. Thus, we eliminate terms with seven or more words, removing over a quarter of UMLS terms.
3. *Maximum number of characters.* Given the histogram in figure 2, only 39 terms have 56 or more characters. We thus eliminate terms with fewer than 2 characters or more than 55 characters, removing over a fifth of UMLS terms.

**Table 3** Top source vocabularies and their degree of utilisation, by number of unique term strings in clinical notes

| (A) UMLS | | (B) Clinical text—Mayo | | | | (C) Biomedical text—Medline | | |
|---|---|---|---|---|---|---|---|---|
| Source | Unique | Source | Unique | % Use | Frequency | Source | Unique | % Use |
| SNOMED-CT | 988 733 | CHV | 106 426 | 74.4 | 1 866 925 442 | MSH | 242 462 | 32.6 |
| MSH | 743 332 | SNOMED-CT | 94 788 | 9.6 | 1 538 745 839 | SNOMED-CT | 215 217 | 21.8 |
| MEDCIN | 726 724 | MSH | 51 584 | 6.9 | 753 847 562 | NCI | 101 807 | 58.0 |
| NCBI | 662 674 | NCI | 50 536 | 28.8 | 981 062 417 | CHV | 85 473 | 59.7 |
| RXNORM | 455 466 | RCD | 42 668 | 12.3 | 1 683 517 327 | NCBI | 84 129 | 12.7 |
| RCD | 346 922 | MEDCIN | 32 335 | 4.4 | 298 650 586 | RCD | 69 519 | 20.0 |
| LNC | 313 431 | SNMI | 30 280 | 18.5 | 629 881 044 | SNMI | 57 177 | 34.8 |
| ICD10 | 249 863 | MDR | 28 714 | 39.8 | 310 815 333 | SCTSPA | 56 735 | 3.8 |
| NCI | 175 679 | MTH | 21 642 | 15.3 | 866 386 287 | OMIM | 46 339 | 34.5 |
| SNMI | 164 069 | SCTSPA | 17 661 | 1.2 | 369 476 316 | MTH | 43 029 | 30.5 |

Frequency of terms from each source in clinical text is also shown.

CHV, Consumer Health Vocabulary; ICD10, International Classification of Diseases, 10th revision; LNC, Logical Observation Identifier Names and Codes (LOINC); MDR, Medical Dictionary for Regulatory Activities Terminology (MedDRA); MSH, Medical Subject Headings; MTH, UMLS Metathesaurus; NCBI, National Center for Biotechnology Information; NCI, NCI Thesaurus; OMIM, Online Mendelian Inheritance in Man; RCD, Clinical Terms Version 3 (Read Codes); SCTSPA, SNOMED Terminos Clinicos; SNMI, SNOMED International v3.5; SNOMED-CT, Systematized Nomenclature of Medicined - Clinical Terms.

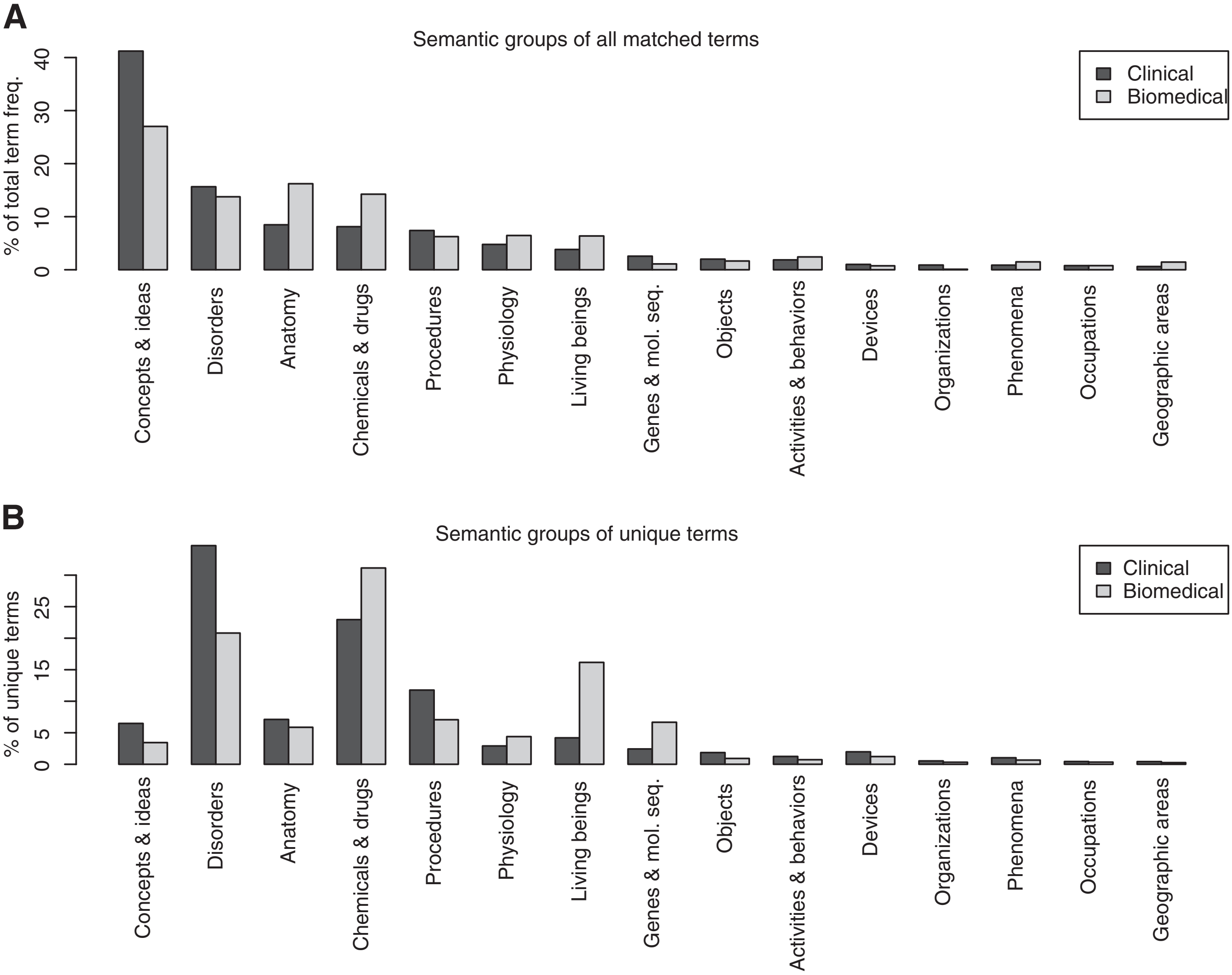

**Figure 5** (A) Frequencies of terms discovered in clinical versus biomedical text, by semantic group; (B) number of unique terms, by semantic group.

4. *Language.* Fifteen languages are represented in the UMLS. Filtering to English terms reduces the set of UMLS terms by almost a third.
5. *Source terminology.* Many UMLS source terminologies are not designed to be lexicons (eg, International Classification of Diseases, ninth revision billing codes). We keep only the top 14 English sources out of the possible 155: SNOMED-CT, Consumer Health Vocabulary, National Cancer Institute (NCI) Thesaurus, Medical Subject Headings (MSH), Read Codes, Medical Dictionary for Regulatory Activities Terminology (MedDRA), SNOMED International, MEDCIN, UMLS Metathesaurus, National Drug File—Reference Terminology (NDF-RT), the original SNOMED, Online Mendelian Inheritance in Man (OMIM), Logical Observation Identifiers Names and Codes (LOINC) and Computer Retrieval of Information on Scientific Projects (CRISP) Thesaurus.
6. *Semantic group.* Of the 15 semantic groups, over 92% of Mayo Clinic terms come from only 7: anatomy, chemicals & drugs, concepts & ideas, disorders, living beings, physiology, and procedures.
7. *Empirical occurrence filter.* We filter out those terms that never appeared in the Mayo corpus. This leaves the full set of Mayo Clinic term occurrences and tests the transferability of a specific lexicon across institutions.
8. *Term frequency.* A total of 99.99% of mentions can be retained if we eliminate terms that occurred only once or twice in the Mayo corpus. This is a subset of the empirical occurrences filter, since zero occurrences are also eliminated.

### Cross-institutional filtering evaluation

Table 4 reports the impact of this filtering. First, we begin with a baseline of the full UMLS. The top left cell indicates the number of unique UMLS terms. Rows show the lexicon size reduction effect of individual filters against this baseline. The final rows apply multiple filters at once.

The left 'UMLS' columns analyze how much of the UMLS Metathesaurus remains after each of the filters, and larger percent reduction values correspond to more memory-efficient systems. The middle 'Mayo' columns evaluate the reasoning for choosing these filter definitions. For example, our semantic groups filter (filter 6 in table 4) uses only seven semantic groups. Reading the row from left to right, it reduces the size of the lexicon to 7 798 937 (a 6.43% reduction), keeps 273 300 of the 296 798 unique terms (ie, excludes 7.92%), and keeps $2.289\times10^9$ of the $2.376\times10^9$ term occurrences (ie, excludes 3.68%) for the Mayo corpus. As a whole, the filters defined in this example might be reasonable for some information extraction applications, excluding only 5.57% of all mentions.

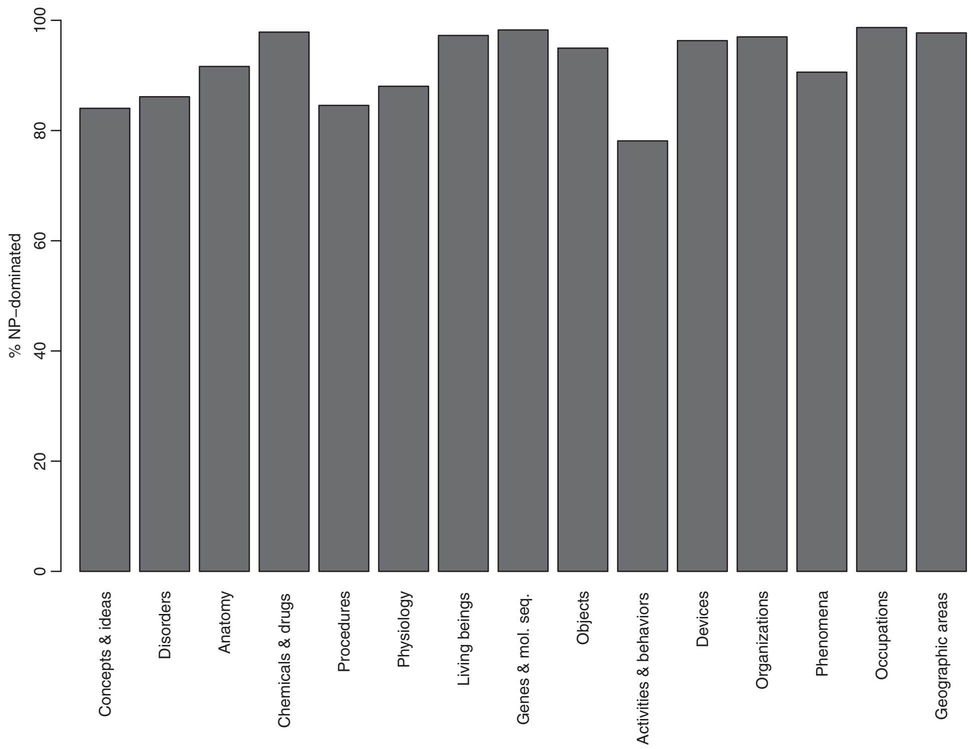

**Figure 6** Percentage of unique terms that are noun phrase (NP) dominated, by semantic group.

The right 'i2b2/VA' columns are defined by using Mayo-based filters on term matches from the i2b2/VA corpus.

Our cross-institutional evaluation lies in comparing the 'Mayo' columns with the 'i2b2/VA' columns. Filters 1—4 seem to apply similarly and accurately across the two corpora. This is to be expected because they largely deal with systematic intrinsic properties of the term strings in the UMLS and should not depend on corpora. The remaining filters differ between Mayo and i2b2/VA data, indicating that statistical analysis along those lines should only be transferred across data sources with caution.

The source terminology filter removed far less a proportion of unique terms in the i2b2/VA corpus (6.14%) than in the Mayo corpus (15.31%). This is probably due to the vast size difference between the two corpora: recalling figure 3, a heavy tail distribution within large corpora means that many uncommon terms are mapped in the Mayo Corpus, but not in the i2b2/VA corpus. We may conclude that filtering by source reduces the diversity of available terms, but the most frequent terms are captured in a small number of sources.

In i2b2/VA data, the semantic group filter excludes a higher proportion of unique terms but a smaller proportion of term mentions than in Mayo data. The variability is not great, however, compared with the differences with the biomedical literature domain in figure 5A,B. We conclude that though different clinical corpora may have slightly different distributions, their utilised terms are still relatively similar to each other in semantic groups.

Perhaps most instructive are filters 7—8, the empirical occurrences and the term frequency filters. Both of these filters exclude a smaller proportion of unique terms in the i2b2/VA data (1.39% for filter 8) than in the Mayo data (23.62% for filter 8), again likely due to the corpus size differential. However, a larger proportion of i2b2/VA mentions are excluded (15.23% for filter 8) than Mayo mentions (0% for filter 8). Despite the fact that these are both clinical corpora, term frequencies have vastly different characteristics in the different corpora. Although these statistics are standard NLP techniques, they would appear to be more helpful within an institution than across sources of data.

**Table 4** Transferability of corpus-based filtering of the Unified Medical Language System (UMLS)

| | UMLS | | Mayo Clinic term occurrences | | | | i2b2/VA term occurrences | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Unique | % rdn | Unique | % exc | Matches (n) | % exc | Unique | % exc | Matches (n) | % exc |
| Full UMLS | 8 335 125 | — | 296 798 | — | $2.376\times10^9$ | — | 17 570 | — | 376 350 | — |
| 1. Sp. Char. | 5 146 096 | 38.26 | 296 798 | 0.00 | $2.376\times10^9$ | 0.00 | 17 570 | 0.00 | 376 350 | 0.00 |
| 2. MaxWord | 6 157 283 | 26.13 | 295 385 | 0.48 | $2.376\times10^9$ | 0.00 | 17 564 | 0.03 | 376 343 | 0.00 |
| 3. MaxChar | 6 477 250 | 22.29 | 296 516 | 0.10 | $2.376\times10^9$ | 0.00 | 17 569 | 0.01 | 376 349 | 0.00 |
| 4. Language | 5 610 576 | 32.69 | 296 167 | 0.21 | $2.375\times10^9$ | 0.05 | 17 552 | 0.10 | 376 234 | 0.03 |
| 5. Sources | 3 409 183 | 59.10 | 251 361 | 15.31 | $2.327\times10^9$ | 2.08 | 16 491 | 6.14 | 368 682 | 2.04 |
| 6. SemGroup | 7 798 937 | 6.43 | 273 300 | 7.92 | $2.289\times10^9$ | 3.68 | 16 343 | 6.98 | 361 018 | 4.07 |
| 7. EmpFilt | 296 798 | 96.44 | 296 798 | 3.56 | $2.376\times10^9$ | 0.00 | 17 371 | 1.13 | 319 258 | 15.17 |
| 8. TermFreq | 230 011 | 97.24 | 226 697 | 23.62 | $2.376\times10^9$ | 0.00 | 17 326 | 1.39 | 319 039 | 15.23 |
| Filters 1—8 | 181 523 | 97.82 | 181 523 | 38.84 | $2.244\times10^9$ | 5.57 | 15 139 | 13.84 | 301 473 | 19.90 |
| Filters 1—6 | 1 448 811 | 82.62 | 230 860 | 22.22 | $2.244\times10^9$ | 5.56 | 15 343 | 12.68 | 354 274 | 5.87 |
| Filters 1—5 | 1 594 674 | 80.87 | 250 192 | 15.70 | $2.327\times10^9$ | 2.09 | 16 486 | 6.17 | 368 676 | 2.04 |

The UMLS column shows % rdn (reduction) of lexicon size (larger % rdn is more efficient). The Mayo and i2b2/VA columns compare this to % exc (exclusion) rate, wherein UMLS terms are no longer mapped due to the filtering. Incongruencies in % exclusion indicate corpus differences.

## DISCUSSION

The foregoing cross-institutional test aligns with envisioned applications because a user of a practical application like Lexicon Builder[4] will need guidance on what filters to choose. It would be safe in such a situation to apply filters that have been validated across institutions, but other filters should be applied with caution. A final recommendation for use-case agnostic information extraction in the i2b2/VA corpus, as presented in table 4, might be to utilise filters 1—5. These simple filters achieve five-fold reduction in lexicon size (efficiency) while preserving almost 94% of unique terms and almost 98% of mentions in the corpus.

The semantic group filter has limited utility because it does not greatly reduce the dictionary size. However, other factors, such as the limitations of a corresponding human annotation effort, may be reasons for narrowing the scope of general information extraction to specific semantic groups.

Most importantly, the results show that the empirical occurrences and term frequency filters are highly institution specific. Any methodologies developed off of these statistics should take care to complete a preliminary corpus analysis rather than directly using the Mayo Clinic statistics.

Unlike our previous work,[12] the preceding analysis does not attempt to calculate or analyse concept-level semantics. Although mixing the two analyses is an interesting problem, our term-level analysis is natural for the envisioned problem setting, where a user is building a lexicon of strings for concept indexing—concept normalisation would presumably be a downstream task. Additionally, we do not calculate the 'usefulness' of filters in real-world applications because such measures typically require a concept-centric focus.

## CONCLUSION AND FUTURE WORK

Based on the occurrences of terms in a 51 million document corpus of Mayo Clinic clinical notes, this paper has presented a suite of statistics on UMLS term occurrences in the clinical domain, and has evaluated the cross-institutional applicability of these statistics. We have shown several measures that are intrinsic to their Metathesaurus entries (term well-formedness, length and language) that generalise easily across clinical institutions. Term frequencies are highly variable across institutions and should be adapted across domains or institutions with caution. The semantic groups of mapped terms may differ slightly from institution to institution, but the distance between institutions is much smaller than that between the clinical and biomedical literature domains.

We believe this analysis makes it possible for end users to build customised, empirically informed lexicons from the UMLS. Implementationally, this team plans on enhancing Lexicon Builder[4] with the statistics presented above. Other future work includes the further characterisation of clinical note sections (eg, terms may differ in history of present illness vs discharge diagnosis sections), types of notes (eg, discharge summaries vs operative reports), co-occurrence information (ie, utilising latent semantic information), and ontological structure (eg, which branches in an ontology are more useful).

As mentioned, a concept-centric analysis and its relationship to our term-centric analysis are also areas of future work. A concept-centric filtering evaluation, for example, may actually show that precision could be improved by filtering, since it could remove 'distracting' terms.

While the coverage of lexicons derived out of biomedical ontologies is impressive, clinical writing contains many more variants. We plan to generate accurate variants by analysing lexical variants, synonyms and related terms at a large scale.

**Acknowledgments** The authors would like to acknowledge Rong Xu, Vinod Kaggal and Yipei Liu for their help on the experiments and statistics, and the anonymous reviewers for their thorough feedback.

**Contributors** SW carried out the experiments, led the study design and analysis and drafted the manuscript. HL and DL helped with coding the experiments and with manuscript drafting. NS and CT enabled the comparisons with biomedical data, and NS also helped draft the manuscript. MM and CC provided institutional support and manuscript editing.

**Funding** This work was supported in part by the NIH Roadmap Grant U54 HG004028. This study was also supported by National Science Foundation ABI:0845523, National Institute of Health R01LM009959A1, and the SHARPn (Strategic Health IT Advanced Research Projects) Area 4: Secondary Use of EHR Data Cooperative Agreement from the HHS Office of the National Coordinator, Washington, DC, DHHS 90TR000201.

**Competing interests** None.

**Ethics approval** Ethics approval was provided by Mayo Clinic Institutional Review Board.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** The aggregate corpus statistics in this paper will be released open-source at a future date as a part of the Lexicon Builder web service.

## REFERENCES

1. **Lindberg DA,** Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993;**32**:281.
2. **Aronson AR,** Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**:229—36.
3. **Savova GK,** Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507—13.
4. **Parai GK,** Jonquet C, Xu R, *et al*. The Lexicon Builder Web service: building custom lexicons from two hundred biomedical ontologies. *AMIA Annu Symp Proc* 2010;**2010**:587—91.
5. **Aho AV,** Corasick MJ. Efficient string matching: an aid to bibliographic search. *Commun ACM* 1975;**18**:333—40.
6. **Dai M,** Shah NH, Xuan W, *et al*. *An efficient solution for mapping free text to ontology terms*. AMIA Summit on Translational Bioinformatics, San Francisco, CA, 2008.
7. **Savova G,** Kipper-Schuler K, Buntrock J, *et al*. *UIMA-Based Clinical Information Extraction System. Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*. Proceedings paper, LREC (Languages Resources and Evaluation Conference), Marrakech, Morocco, 2008:39.
8. **Uzuner O,** South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;**18**:552—6.
9. **Aronson A.** Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001:17—21.
10. **Nadkarni P,** Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. *J Am Med Inform Assoc* 2001;**8**:80—91.
11. **Denny JC,** Smithers JD, Miller RA, *et al*. 'Understanding' medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc* 2003;**10**:351—62.
12. **Wu S,** Liu H. Semantic characteristics of NLP-extracted concepts in clinical notes vs. biomedical literature. *Annual Symposium of American Medical Informatics Association*, Washington DC, WA, 2011.
13. **Aronson AR.** Filtering the UMLS Metathesaurus for MetaMap. *National Library of Medicine Technical Report*. 2006.
14. **Halevy A,** Norvig P, Pereira F. The unreasonable effectiveness of data. *Intelligent Systems IEEE* 2009;**24**:8—12.
15. **McCray AT,** Bodenreider O, Malley JD, *et al*. Evaluating UMLS strings for natural language processing. *Proc AMIA Symp* 2001:448.
16. **Hettne KM,** van Mulligen EM, Schuemie MJ, *et al*. Rewriting and suppressing UMLS terms for improved biomedical term identification. *J Biomed Semantics* 2010;**1**:5.
17. **Thompson P,** McNaught J, Montemagni S, *et al*. The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics* 2011;**12**:397.
18. **Lippincott T,** Seaghdha D, Sun L, *et al*. Exploring variations across biomedical subdomains. *Proceedings of International Conference on Computational Linguistics*, Beijing, China, 2010:689—97.
19. **Cohen KB,** Johnson HL, Verspoor K, *et al*. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics* 2010;**11**:492.
20. **Xu R,** Musen MA, Shah NH. A comprehensive analysis of five million UMLS metathesaurus terms using eighteen million MEDLINE citations. *AMIA Annu Symp Proc* 2010;**2010**:907—11.
21. **Michel JB,** Shen YK, Aiden AP, *et al*. Quantitative analysis of culture using millions of digitized books. *Science* 2011;**331**:176—82.
22. **Chute CG,** Beck SA, Fisk TB, *et al*. The enterprise data trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. *J Am Med Inform Assoc* 2010;**17**:131.
23. **Bodenreider O,** McCray AT. Exploring semantic groups through visual approaches. *J Biomed Inform* 2003;**36**:414—32.